# Modeling the Past, Present, and Future of Influenza

David Farrow

CMU-CB-16-101

July, 2016

Computational Biology Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

**Thesis Committee:**
Roni Rosenfeld, Chair
Ryan Tibshirani
Carl Kingsford
John Grefenstette
Elodie Ghedin

*Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.*



This work was supported by NIH T32 training grant T32 EB009403 as part of the HHMI-NIBIB Interfaces Initiative; and by the National Institute of General Medical Science of the NIH under award number U54GM088491.

The views expressed herein are solely the responsibility of the author and do not necessarily represent the official views of the National Institutes of Health or any other entity.

**Keywords:** influenza, viral evolution, phylodynamics, epidemiological forecasting, epidemiological nowcasting, sensor fusion

*To my family and friends, for your unending inspiration and support.*

# Abstract

Influenza has been, and continues to be, a significant source of disease burden worldwide. Regular epidemics and sporadic pandemics are incredibly costly to society, not just in terms of the monetary expense of prevention and treatment, but also in terms of reduced productivity, increased absenteeism, and excessive morbidity and mortality. Major obstacles to mitigating these costs include an incomplete understanding of influenza's phylodynamics, the inherent delays of clinical surveillance and reporting, and a lack of outbreak forewarning.

The aim of this thesis is to address each of these obstacles computationally by (1) simulating transmission and evolution of influenza to explore the interplay between human immunity and viral evolution; (2) collecting and integrating a diverse set of real-time digital surveillance signals to track influenza activity; and (3) generating season-wide forecasts of influenza epidemics using an ensemble of statistical models, simulations, and human judgment.

The first part explores the concept of generalized immunity, which was previously hypothesized to be highly protective but short-lasting. Large-scale, long-term simulations based on an extension of an earlier model were used to scan immunity parameter space and indicate that the most plausible definition of generalized immunity is less protective but potentially much longer-lasting than previously assumed. The second part describes how sensor fusion and tracking can be applied to the nowcasting problem. Drawing from control theory, weather forecasting, and econometrics, an optimal filtering methodology is developed to integrate a set of proxies for influenza activity which share one common property: they are available online and in real-time. Otherwise, they are available at different temporal intervals, geographic resolutions, and historical periods, and they are noisy and potentially correlated. The resulting nowcasts are robust to failure of individual proxies and are available up to several weeks before traditional surveillance reports. The third part combines earlier results with novel methodologies to produce probabilistic forecasts of influenza spread and intensity that are timely, accurate, and actionable. In particular, an empirical Bayes method and spline regression are used to produce forecasts which only rely on the availability of historical data and are readily generalizable to other infectious diseases; and a wisdom of crowds approach is used to incorporate human judgment into the forecasting process.

# Acknowledgments

I am, of course, greatly indebted to a large number of people, without whom this thesis would not have been possible. It is difficult to adequately express just how significant the contributions of friends, family, colleagues, and mentors have been, not only in my graduate career, but more generally in the grand scheme of life.

My first thanks go to my wonderful wife, Brittany, for being with me every step of the way. This thesis is as much due to her loving and unwavering support as it is to my own effort. Were it possible to share a PhD, she would be due a share of the title no smaller than my own. I dare not imagine how things would have turned out without her by my side.

This thesis has been shaped positively and fundamentally by the patient and kind guidance of my graduate advisor, Roni. From the first day of rotation in his lab, I have felt incredibly fortunate to work for and with such an intellectual giant; I feel even more fortunate today as I write these last few words in my thesis. Roni, it seems, has the inexplicable ability to push me to heights I never knew were possible to achieve, and without him the contributions of this thesis would have fallen much further short of what I was unknowingly capable. I must also say that the friendship I have shared with Roni over the years has been as important to my successes as anything else.

There are many other researchers and professors who deserve ample credit for their varied contributions along the way. My committee has been critical in shaping the overall direction of my work, and the feedback I received along this journey from each of Ryan, Carl, John, and Elodie has been of tremendous value. Likely unbeknownst to each of them, I have gathered an incredible amount of knowledge and skill from our varied interactions over the years, and this has certainly been crucial in my academic and professional development.

I would additionally like to express my gratitude to a number of organizations and individuals who, in a very practical sense, enabled the work herein. Several United States government agencies have hosted, and continue to host, epidemiological forecasting challenges and workshops, including the Centers for Disease Control and Prevention (CDC), Defense Advanced Research Projects Agency, and The White House Office of Science and Technology Policy. Data, tools, and collaboration were provided by Google, the Johns Hopkins University Social Media and Health Research group (SMHR), the Wikimedia Foundation, and CDC. I would like to thank in particular: Christian Stefansen, Shlomo Urbach, and Lisa Creed (Google); Mark Dredze (SMHR); and Matthew Biggerstaff, Michael Johansson, and Lyn Finelli (CDC).

There is no doubt that I was set on this trajectory early in life, and for this I am eternally grateful to my loving and supportive parents, Carl and Allison. Were it not

for their nurture and encouragement, I most certainly would not have made it this far. Because of them, I chased my dreams. My dear sister, Alyssa, deserves much credit in this regard as well. To my extended family, including Barry and Beth, I similarly give my gratitude for support and encouragement over the years.

My sanity throughout this endeavor was maintained in no small part by the efforts of my good friends Sean, Bradley, and Lance. For the adventures, the advice, the perfect mixture of distraction and motivation, and the far too infrequent late nights of caffeine, code, and dota, I am grateful.

Finally, I would be remiss if I did not at least mention the giants throughout my life who in very significant ways helped me get to this point. My sincerest thanks go out to: Dennis Cabaniss and Julie Goode, for introducing me to computer science and for supporting my development through participation in the University Interscholastic League; and to Tim Baird, Scott Ragsdale, Steve Baber, Steve Moore, Rebekah Rampey, and Kevin Stewart, for teaching, motivation, encouragement, and being superb scientists and role models.

# Contents


**1 Introduction** **1**
- 1.1 Motivation . . . 1
- 1.2 Background . . . 2
  - 1.2.1 Virology and Pathogenicity . . . 2
  - 1.2.2 Epidemiological Surveillance . . . 3
- 1.3 Overview . . . 4
  - 1.3.1 Thesis statement . . . 4
  - 1.3.2 Scope . . . 4
  - 1.3.3 Approach . . . 5

**2 Related Work and New Directions** **7**
- 2.1 Recapitulation . . . 7
- 2.2 Innovation . . . 11

**3 Inferring Parameters of Human Immunity by Modeling Influenza** **13**
- 3.1 The evolutionary conundrum . . . 13
- 3.2 A working model of influenza . . . 14
  - 3.2.1 Description of the base model . . . 14
  - 3.2.2 Questioning the mechanisms of immunity . . . 15
  - 3.2.3 Model implementation and extensions . . . 16
- 3.3 Assessing outcomes . . . 17
  - 3.3.1 Epidemiological features . . . 17
  - 3.3.2 Evolutionary features . . . 17
- 3.4 Mapping the parameter space of generalized immunity . . . 18
- 3.5 Computing likelihood across parameter space . . . 24
- 3.6 Sensitivity analysis and robustness . . . 26
  - 3.6.1 Relaxing assumptions . . . 26
  - 3.6.2 Targeting median age instead of life expectancy . . . 28
  - 3.6.3 A more realistic population structure . . . 29
- 3.7 Final considerations . . . 30

**4 Nowcasting Influenza through Sensor Fusion of Digital Surveillance** **33**
- 4.1 Situational awareness for preparedness . . . 33
  - 4.1.1 The gold standard is not ground truth . . . 34

4.1.2 The rise of digital surveillance . . . 35
4.2 A strategy for optimal assimilation . . . 36
4.2.1 The Kalman filter . . . 36
4.2.2 Derivation of the sensor fusion kernel . . . 37
4.3 Proxies of flu activity in the US . . . 41
4.3.1 Measurements . . . 41
4.3.2 Predictions . . . 45
4.3.3 Summary . . . 47
4.3.4 Fitting digital surveillance to (w)ILI . . . 48
4.4 Nowcasting influenza within the US . . . 51
4.4.1 Digital and predictive surveillance in the sensor fusion framework . . . 51
4.4.2 Results and comparative analysis . . . 53
4.4.3 Sensitivity analysis . . . 57
4.5 Final considerations . . . 60

**5 Forecasting Influenza Epidemics using Statistical Models and Human Judgment 63**
5.1 Learning from the past . . . 63
5.2 An intuitive approach . . . 64
5.2.1 Predictions, forecasts, and accuracy . . . 64
5.2.2 A first attempt . . . 65
5.3 Frameworks for epidemiological forecasting . . . 66
5.3.1 The *Empirical Bayes* forecasting framework . . . 66
5.3.2 The *Epicast* forecasting framework . . . 70
5.4 Results in forecasting the 2014–2015 flu season . . . 72
5.4.1 Objectives, Targets, and Accuracy . . . 72
5.4.2 Epicast Participation and Standalone Accuracy . . . 74
5.4.3 Comparison of Accuracy Between Forecasting Methods . . . 78
5.4.4 Results in forecasting Onset Week . . . 81
5.5 Adaptive extensions to the Epicast framework . . . 82
5.5.1 The relative accuracy of expert and non-expert predictions . . . 83
5.5.2 A weighting scheme based on expertise and past performance . . . 85
5.6 Epidemiological forecasting of other diseases . . . 88
5.6.1 OSTP Dengue Challenge . . . 88
5.6.2 DARPA Chikungunya Challenge . . . 90
5.7 Final considerations . . . 93
5.7.1 Pros, Cons, and Caveats of human judgment in forecasting . . . 93
5.7.2 The state of the art . . . 94
5.7.3 A model of models . . . 94

**6 Conclusion 97**
6.1 Summary of contributions . . . 97
6.2 Future directions . . . 99
6.3 Final thoughts . . . 100

**Appendix A The *Pinned Spline* forecasting framework** **103**
A.1 Intuition . . . . . 103
A.2 Regression Splines . . . . . 104
A.3 From point prediction to distributional forecast . . . . . 105

**Appendix B The *Archefilter* forecasting framework** **107**
B.1 Overview . . . . . 107
B.2 The Archetype . . . . . 107
B.3 The Archefilter . . . . . 109
B.4 Use within the Kalman filter . . . . . 112

**Bibliography** **115**

# List of Figures


3.1 Representative phylogenies of RNA viruses . . . 14
3.2 Epidemiological parsimony across immunity parameter space . . . 20
3.3 Evolutionary parsimony across immunity parameter space . . . 21
3.4 Characteristic dynamics of various parameter regimes . . . 23
3.5 Plausible parameterizations of generalized immunity . . . 25
3.6 Plausible parameterizations under varied assumptions . . . 27
3.7 Average simulated correlation structure . . . 28
3.8 Plausible parameterizations using median age . . . 29
3.9 Plausible parameterizations including tropical deme . . . 30

4.1 Significant wILI adjustment due to backfill . . . 35
4.2 Google Trends as a digital surveillance signal . . . 36
4.3 Regression weights for fitting signals to (w)ILI . . . 49
4.4 Plot of sensor readings for the US . . . 50
4.5 Overview of the sensor fusion and extraction process . . . 52
4.6 Retrospective nowcasts for all US locations . . . 54
4.7 Comparison of sensor fusion inputs and output . . . 56
4.8 Sensor fusion ablation experiments . . . 58

5.1 Time series of wILI in the US . . . 64
5.2 Trendfiltered wILI trajectories . . . 67
5.3 Transformations of wILI trajectories . . . 68
5.4 Epicast user interface . . . 71
5.5 Backfill causes a two week shift of Onset Week . . . 74
5.6 Overview of 2014–2015 Epicast participation . . . 75
5.7 Overall accuracy of Epicast for short-term targets . . . 76
5.8 Accuracy of Epicast by lead time for all targets . . . 77
5.9 Mean absolute error of Epicast by lead time for all targets . . . 78
5.10 Epicast Win Rate against individual human predictions and competing systems . . . 79
5.11 Comparison of log scores for Epicast and Empirical Bayes . . . 80
5.12 Accuracy in forecasting Onset Week . . . 82
5.13 Expert versus Non-expert MAE . . . 83
5.14 Expert versus Non-expert MLL . . . 84
5.15 Expert versus Non-expert Win Rate . . . 85
5.16 User Weight over Time . . . 87

5.17 MAE of Weighted Epicast . . . . . 87
5.18 Dengue challenge results . . . . . 90
5.19 Overview of Epicast chikungunya forecasts . . . . . 92

A.1 Spline basis and B-spline fit . . . . . 104

B.1 Graphical representation of the influenza archetype . . . . . 108
B.2 Archetpye transformations and best fit . . . . . 110
B.3 Parameter space of archetypes fit to wILI . . . . . 111

# List of Tables


3.1 Summary of epidemiological and evolutionary measures . . . . . 19

4.1 Summary of digital surveillance and forecasting signals . . . . . 48
4.2 Nowcasting accuracy by location . . . . . 55
4.3 Comparison of nowcast and preliminary wILI . . . . . 55
4.4 Comparison of nowcasting frameworks . . . . . 57
4.5 Sensor fusion abscission experiments with all sources . . . . . 59
4.6 Sensor fusion abscission experiments with selected sources . . . . . 59

5.1 Mean absolute error by forecasting target and system . . . . . 81

# Chapter 1

# Introduction

> We have grown accustomed to the wonders of clean water, indoor plumbing, laser surgery, genetic engineering, artificial joints, replacement body parts, and the much longer lives that accompany them. Yet we should remember that the vast majority of humans ever born died before the age of 10 from an infectious disease.
>
> *Stuart Jay Olshansky*

## 1.1 Motivation

Disease is an unfortunate reality of our biological existence. The burden it places collectively on humanity is tremendous, and the cost of disease—measured in terms of economic impact, disability-adjusted life years, and years of life lost—is staggering. In the 1993 World Development Report, it was estimated that on the order of 40% of the total disease burden is attributable to infectious diseases [1]. This figure is unsettlingly high given the tremendous advances made towards the control and prevention of infectious diseases through the scientific and industrial revolutions.

The effects of such advances are exemplified by the rapid decline in infectious disease mortality in the United States of America (US) during the early 20th century [2]. That such a decline took place is probably unsurprising since improvements in hygiene, education, and technology are associated with a reduction in disease burden [3]. But what *is* surprising is that in spite of continued advances across a variety of fields, the rate of infectious disease mortality in the US has not fallen since around the 1960s—in fact, it has been rising since the 1980s [4].

One of the most ubiquitous infectious diseases in modern society is influenza (flu), the condition resulting from infection by the *Influenzavirus* genera of the *Orthomyxoviridae* virus family. Flu has been, and continues to be, a pestilence on humanity, and it is conservatively estimated that 10% of the global population will be infected—*every year* [5]. Such widespread infection

results in an estimated 250,000 annual flu-related deaths globally [6]. This is due in no small part to the fact that we have, so far, been unable to produce a flu vaccine with broad, effective, and lasting protection—though progress is being made [7]. Current flu vaccines are redesigned each year out of necessity and generally only provide partial and temporary protection, and that only for a small selection of circulating strains [8]. New strains of flu arise with alarming regularity, and the threat of the next global pandemic is ever present, if not imminent [9, 10].

Fortunately, in our present situation, there is room for improvement. Through the recent digital revolution, and its even more recent product, the information revolution, we have acquired a set of complementary and extremely powerful tools: *computation* and *data*. In this thesis I use computer science, machine learning, and statistics to tackle some of the biggest obstacles hindering progress in developing a better understanding of, in preparing for, and in mitigating the effects of flu and other infectious diseases.

## 1.2 Background

### 1.2.1 Virology and Pathogenicity

The virology of influenza is fascinating, having a certain quality of deadly elegance. I summarize the salient points here, but I refer the interested reader to the book *Fields Virology* for a more thorough treatment of the subject [11].

Influenza is a segmented, enveloped virus, whose genetic information is stored in the form of ribonucleic acid (RNA). There are several *types* of influenza, and of these, humans are a natural reservoir for types A and B. These two types share many common properties, but all known pandemics to date, and the worst seasonal epidemics, are caused by type A, and so for the remainder of this thesis I focus primarily on this type. The 13.5 kilobase genome of Influenza A virus (IAV) is made up of eight distinct segments of RNA, each encoding one or more proteins required for infection and replication. The most important segments for this thesis are those which encode the proteins Hemagglutinin (HA) and Neuraminidase (NA). Both of these proteins are present on the surface of the virion, and together they provide a molecular signature that is recognizable by the human immune system.

In humans, influenza infects epithelial cells of the upper and lower respiratory system, most often causing symptoms of fever and cough. In fact these symptoms (fever of at least 100 °F with cough and/or sore throat, without a known cause other than influenza) are what is defined by the US Centers for Disease Control and Prevention (CDC) and by the World Health Organization (WHO) as influenza-like illness (ILI) [12, 13].

Influenza has several properties that make it a particularly insidious pathogen. Foremost among these properties is its RNA—as opposed to deoxyribonucleic acid (DNA)—genome. To understand why this is of concern, it is necessary to understand a little bit about the biological process of genomic replication. At the most basic level, the same molecular machinery is required to replicate either form of nucleic acid. As with all biological processes, mistakes, however rare, are possible and do occasionally occur. Fortunately for DNA-based lifeforms, like us, the machinery for replicating DNA is quite sophisticated, containing stringent quality control checks and error-correcting capabilities. This level of sophistication is generally absent in the

more primitive RNA-based lifeforms, including influenza and many other viruses. The result is that errors in replication, called mutations, accumulate much more rapidly in RNA-based life (roughly $10^{-5}$ mutations per nucleotide per replication) than in DNA-based life (roughly $10^{-9}$ mutations per nucleotide per replication) [14]. Although mutations are usually deleterious, they may occasionally be beneficial. For the case in point, mutations in the influenza genome could potentially obscure the molecular signature of the virus, hindering recognition by a host's immune system. In this context, the process of continual mutation over time is known as antigenic drift, and it is the source from which new strains of influenza continually arise. More generally, the gradual accumulation of mutations throughout the entire genome (as opposed to just regions encoding epitopes) is termed genetic drift.

Another property of influenza is that its genome is segmented. In the event of coinfection—a cell being infected with two or more different strains—it is possible that new virions could emerge containing a genetic mixture of the infecting strains. This process is called reassortment, and it is the way in which pandemic strains arise. A novel influenza virus carrying a mixture of genes from a distinct set of parent strains has the potential to be almost completely unrecognizable to the immune system—an antigenic shift.

Finally, influenza is capable of direct airborne transmission—even before the first symptoms appear [15]. As a result, influenza can be extremely contagious and can rapidly spread throughout a population. With the relatively recent advent of human air travel, and given the increasingly interconnected global population, influenza has the potential now to spread throughout the world in record time. For these reasons, influenza—in both epidemic and pandemic forms—is a persistent and increasingly serious threat, globally.

### 1.2.2 Epidemiological Surveillance

For over a century, records have been kept that describe influenza case counts and mortality in the US. Historically, this data was collected and distributed by independently acting government and health departments. As a result of differences in reporting practices and case definitions, this data can be difficult to aggregate on a national level. Although the case definitions have been refined and new technologies allow for rapid and accurate diagnosis, there are still differences in reporting across the US. This is due at least in part to the fact that influenza is in general not a nationally notifiable disease—although some specific forms, such as novel antigenic types and cases of pediatric mortality, must be reported. There has been, however, a push in the last two decades to standardize and incentivize influenza reporting through the US Outpatient Influenza-like Illness Surveillance Network (ILINet).

Started in 1997, ILINet is a system run by the National Center for Immunization and Respiratory Diseases (NCIRD) branch of the CDC which collects influenza data on a volunteer basis from health care providers throughout the country. Each week, these providers report, among other things, the total number of patients seen and the fraction of those which had symptoms of ILI. The percent of ILI cases out of total cases gives a number called %ILI (or, simply "ILI"; although this usage is somewhat ambiguous as "ILI" is also the acronym of the clinical definition). CDC collects these reports from all states and publishes %ILI aggregated across broad regions of the country [16]. Because participation in ILINet varies greatly between states, CDC normalizes the data when aggregating across regions; the result is called weighted %ILI (%wILI, or sim-

ply "wILI"). wILI—the percent of ILI cases out of total cases, weighted by state population—is made available in the nine census regions, the ten Health and Human Services (HHS) regions, and for the US as a whole.

Participation in ILINet, which is entirely voluntary, was relatively low in the first several years. As a result, these first seasons are especially noisy (particularly in the smaller regions), and data is missing for weeks outside of the official flu season. In the US, epidemiological weeks, which I refer to as "epiweeks", traditionally follow the Morbidity and Mortality Weekly Report (MMWR) week numbering definition [17]—weeks start on Sunday, and the week containing January 4th is the first week of the year. The flu season is officially defined as epiweeks 40 through 20, roughly spanning October to May of adjacent calendar years. The "on-season" describes this set of weeks, and CDC reports wILI during the on-season throughout all the years of ILINet's existence (since 1997). The "off-season" (alternatively, the "pre-season") is defined by the remaining weeks; epiweeks 21–39, or roughly June through September. As previously mentioned, wILI was initially unavailable during the off-season; however, starting in 2003, and continuing through the present, wILI is available on all weeks of the year.

While ILINet represents a significant improvement in the surveillance of influenza, there are some caveats that should be mentioned. First, there is an inherent one, and sometimes two, week delay between patient care and publication of wILI. This is to be expected because it takes time and human effort to collect and report these statistics—and this is on a volunteer basis. Second, previously published values of wILI are revised over time as new reports are gathered—a process known as "backfill". This, too, is to be expected; providers are asked to report their case counts within a few days, but understandably, delays arise from time to time. As a result, wILI on any given week is subject to change in reports published on subsequent weeks. Third, ILI is a broad syndromic definition which often includes non-influenza infections that cause influenza-like symptoms. Even with these shortcomings, wILI is a very useful signal for epidemiological surveillance and is often considered to be the "gold standard" indicator of flu activity in the US.

## 1.3 Overview

### 1.3.1 Thesis statement

Within this thesis, I explore the following question: what about influenza, and infectious diseases in general, can we learn with what's available *now* to minimize the impact of impending outbreaks? In light of this, my thesis can be summarized thusly:

> *With the computational tools and varied datasets currently available, we can better understand the role of human immunity in shaping viral evolution, estimate disease incidence in real-time, and predict the trajectory of disease outbreaks.*

### 1.3.2 Scope

Influenza is far from alone in causing human disease. While the methods developed in this thesis are inspired by, and specifically attempt to address, the challenges associated with influenza, they are readily generalizable to other viruses, diseases, and domains. As a demonstration of the

general applicability of these methods, I provide case studies for two other diseases: dengue and chikungunya. Like flu, both of these diseases are caused by RNA viruses; but unlike flu, they are spread by mosquitoes, and hence their geographic distribution is somewhat limited.

My methodology—simulation, assimilation, regression, extrapolation, and more—is inherently computational. It follows then that it is also data-dependent. Infectious disease surveillance is not a new development, but it has often been sporadic, coarse-grained, and non-standardized. Flu is a disease for which a we have the rare luxury of having (relatively) historically-rich, high-resolution, and well-defined datasets. Still, each of these alone is far from ideal, and a recurring theme throughout this thesis is to make best use of a combination of any and all available sources of information.

This thesis is primarily a showcase of the work I have done over the last several years, but as in all scientific endeavors, nothing has been done entirely in isolation. I have worked with many mentors, peers, and collaborators, and it is probably impossible to find a project that I completed entirely by myself. There are many projects that I would like to cover in this thesis, but I limit myself only to those for which I have been the primary contributor. Sometimes, however, it is necessary to frame my projects within the context of other projects in which I have been only tangentially involved with. As these cases arise, I make clear the extent of my contribution.

### 1.3.3 Approach

I begin with a broad overview in Chapter 2 of the work related to these topics. This chapter is broken into a review, which serves the dual purpose of initiating the reader and contextualizing the work in this thesis, and an explanation of how subsequent chapters expand on these prior works.

In Chapter 3 I explore the interplay between human immunity and influenza's evolution. Although the exact mechanisms of the immune response against influenza remain unclear, I show that an individual-based model of influenza's transmission and evolution can be used to infer some of the missing details. More specifically, I characterize the likelihood of immunity parameter space, given simulated outcomes and a set of epidemiological and phylogenetic measures for which we have strong empirical evidence.

In Chapter 4 I consider the problem of "nowcasting"—estimating disease incidence in real-time. I begin by describing the problems associated with authoritative datasets and consider an alternative definition of ground truth. Next, I review the available ILI proxies, which consist largely of novel and nontraditional signals collectively referred to as digital surveillance. Starting with a well-known method in control theory—the Kalman filter—I derive an adaptive method of sensor fusion and apply this methodology to nowcast influenza incidence throughout the US.

In Chapter 5 I describe the historical ILI data available for the US and motivate methods for forecasting of this type of data. I demonstrate two forecasting strategies rooted in machine learning and human judgment and show how these systems can be applied to forecast influenza and other infectious diseases. I conclude with a discussion of the state of the art and how the epidemiological forecasting landscape will likely change in the future.

Finally, I conclude in Chapter 6 with a discussion of the contributions in this thesis in the broader context of the scientific enterprise and with an exploration of directions for future work.

# Chapter 2

# Related Work and New Directions

> If I have seen further than others, it is by standing upon the shoulders of giants.
>
> *Isaac Newton*

## 2.1 Recapitulation

Because of the global and recurring nature of influenza outbreaks, there has been a great deal of time and effort devoted to understanding the dynamics of influenza, especially in recent decades. These dynamics can be broadly divided into two categories: evolutionary and epidemiological. I briefly alluded to these topics in Chapter 1, and now I provide a much more thorough exposition of the recent research in these two areas. After this, I explore how an understanding of influenza's dynamics has led to developments in predicting when and where future outbreaks will occur.

Evolutionary dynamics describe how the genetic and antigenic properties of influenza change over time. Interestingly, it appears that some of these dynamics are unique to, and characteristic of, influenza. Perhaps the most prominent example of this is the observation that influenza exhibits remarkably constrained diversity in comparison to other RNA viruses [18]. As a result of this constrained diversity, clades are replaced serially following infrequent, but punctuated, changes in antigenicity [19]. Despite good correlation between genetic and antigenic drift on long timescales [20], antigenic changes are more abrupt than the steady genetic drift on short timescales [21]. Genetic changes have been shown to occur most frequently within the genomic regions that encode the epitopes of the major surface antigen Hemagglutinin (HA) [22, 23], suggesting a strong positive selection for strains able to evade the host immune response. In agreement with epidemiological results, the strains comprising the trunk of the viral phylogeny typically arise in the tropical regions of the world, whereas strains originating in temperate regions often appear only briefly before going extinct, forming short side branches [24, 25, 26]. The distinctive phylogeny of influenza inspired the creation of a new evolutionary measure, kappa ($\kappa$), to quantify and compare the degree of branching within phylogenetic trees [27]. This new measure captures, in a single dimensionless number, one of the most unique aspects of influenza's evolutionary dynamics; among RNA viruses, and even among other types and subtypes of influenza, the A/H3N2 lineage is exceedingly and unexpectedly slender.

In contrast to evolutionary dynamics, epidemiological dynamics are a function of the interaction between influenza and its host—humans in this case. The epidemiological literature includes several quantities which describe various aspects of an outbreak. Incidence, for example, is the number of new cases within any given unit of time. It has been known for quite some time that flu incidence can be approximately recapitulated with compartmental models, for example the susceptible infectious recovered (SIR) models and their extensions [28]. It is also well known that global pandemics occur concomitantly with the introduction of new subtypes [5]. These statements regarding the global dynamics of flu are indicative of a general and high-level understanding of flu dynamics, but they highlight the absence of more specific knowledge of epidemic processes. More recent studies have aimed to sharpen our understanding by elucidating the more fine-grained temporal and spatial patterns of transmission. It has more recently come to light that there is likely a human reservoir within Southeast Asia which appears to be the source of seed strains for annual epidemics around the world [24, 29, 30, 31]. Further, there is a strong correlation between latitude and timing of epidemics, with incidence in temperate regions strongly influenced by both seasonality [32] and absolute humidity [33, 34]. However, rates and patterns of transmission have in general been shown to be more strongly correlated with patterns of human movement than with geographic distances [26, 35, 36, 37]. Despite widespread herd immunity in humans, influenza displays a high reproductive number each season [38], with a non-uniform distribution of incidence over host age. The earliest epidemic onset and the highest attack rates both occur in children [39].

Computational modeling, a relatively recent development, has been used extensively to help refine our understanding of the dynamics of influenza. Mathematical and agent-based models, whether deterministic or stochastic, have been able to capture many of the unique dynamics of influenza, including multiple cocirculating strains [40] and stable oscillations in incidence [41]. Additionally, models have been used to test different biological hypotheses explaining influenza's characteristically limited genetic diversity and linear phylogeny. One such hypothesis suggests the existence of a short-term, strain-transcending immunity (also referred to as non-specific or *generalized immunity*) in humans [42]. To date, several models incorporating generalized immunity have been published, and all are able to generate flu-like outcomes [42, 43, 44, 45, 46]. This particular hypothesis is explored in great detail in Chapter 3.

Here, however, I describe two prominent alternative hypotheses to generalized immunity. The first is known as *epochal evolution* [47]. This hypothesis is rooted in the empirical observation that there is a highly nonlinear—to some extent unpredictable—relationship between genetic distance and antigenic distance. Typical models which define antigenic distance in terms of genetic distance (for example, Hamming distance with bit string genomes) are therefore unable to capture this dynamic relationship. Further, it has been observed that clusters of related strains arise every 2–8 years, replacing the previously dominant cluster and becoming the new dominant cluster. These serial cluster transitions are concomitant with amino acid substitutions that result in punctuated changes in antigenicity. Perhaps surprisingly, the genetic distance between strains in different clusters is almost always smaller than the genetic distance between strains within any single cluster. Keolle, Cobey, et al. designed a model which is capable of generating both the constrained genetic diversity and the periodic cluster transitions expected for influenza. The critical component of this model is a neutral network that maps from genotypes to phenotypes. Under this construction, it is possible to have strains with several amino acid differences map to

the same antigenic state. Similarly, it is possible for strains that differ by only a single amino acid to belong to different antigenic states. Simulated outcomes produced under this model agree qualitatively with empirical outcomes observed for influenza A/H3N2.

Another hypothesis regarding the role of human immunity in constraining the standing genetic and antigenic diversity of influenza is known as *canalized evolution* [25]. This hypothesis is built on the observation that the empirical antigenicity of influenza A/H3N2 over time can be represented roughly, abstractly, and without significant loss of information by points in a plane. In other words, the antigenic trajectory of influenza can be understood in terms of a path through some low-dimensional antigenic space. Bedford at al. built a model wherein the genetic and antigenic state of any given strain is represented by a vector in Euclidean space (nominally $\mathbb{R}^2$). Evolution in this model is represented by updating the position of a strain's point in a direction selected uniformly at random and a distance selected from a gamma distribution. Large-scale individual-based simulations using this model give rise to results expected for influenza, including constrained genetic and antigenic diversity, periodic sweeping cluster transitions, and a linear phylogeny. It was demonstrated that the model, although stochastic, had a tendency for repeatability. Specifically, continuation replicates of a simulation that was frozen in time generally produced similar outcomes on short (1–2 year) timescales. This observation led to the prescient conclusion that the trajectory of influenza may be predictable.

Evolutionary and epidemiological dynamics are used to explain past outbreaks, but there is rapidly growing interest in applying this understanding forward in time to predict the future evolution and transmission of influenza. Predicting the evolutionary dynamics of influenza is critical, for example, in vaccine selection each year. Vaccine strains are selected based on analysis of circulating strains by human assessment of which of those are likely to dominate in subsequent outbreaks [30, 48, 49]. Vaccination, while certainly helpful in preventing the transmission of influenza, is only partially effective [8, 50, 51, 52, 53]. While the complete eradication of influenza in humans is unlikely without a broadly effective and universal vaccine, there are still ways in which the impact of epidemics can be further reduced. To be more specific, what we need now is to improve our capacity for preparedness and prevention—we need forewarning [54, 55]. This is the defining problem for which the nascent field of epidemiological forecasting has risen.

The first attempts at influenza forecasting essentially treated the problem as an instance of more general time series forecasting [56, 57, 58, 59, 60, 61, 62]. As such, many of these ideas, including in particular autoregression and the method of analogues, were borrowed from econometrics and meteorology to predict incidence, emergency department visits, hospitalizations, and mortality due to influenza, pneumonia, and related respiratory infections. The common thread of these approaches is that traditional surveillance is the sole predictor variable. A new and significant development in flu forecasting came with the realization that auxiliary data sources could be integrated to produce better estimates of flu activity. Examples of such sources used in predicting influenza include climatological data [63, 64, 65], search engine queries [66, 67, 68, 69, 70, 71, 72, 73, 74], public comments on social media like Twitter [75, 76, 77, 78, 79, 80], and online information-seeking behavior on websites like Wikipedia [81, 82, 83]. As the state of the art advances, a serious challenge in both nowcasting and forecasting of influenza, and of infectious diseases in general, is to optimally assimilate all of the surveillance signals that are available at runtime.

Predicting the real-time distribution and prevalence of influenza—the *nowcasting* problem—

has traditionally been approached through use of a single digital surveillance stream (for example, Google Flu Trends [69] and HealthTweets [84]). However, much more recent work has focused on assimilating multiple data streams to produce a unified influenza nowcast [85]. Santillana et al. use data streams based on Google searches, Twitter microblogs, Aetna electronic health records, and Flu Near You participatory surveillance. Rather than using a single assimilation method, several approaches rooted in machine learning were used to explore the potential for nowcasting at various lead-times. These methods include stacked linear regression with $\ell^1$ penalty (LASSO), support vector machine regression with either linear or radial basis function kernel (SVM, SVMRBF), and AdaBoost regression with decision trees. These methods were used to assimilate the various digital surveillance streams, producing nowcasts and short-term forecasts each week for the US as a whole. For nowcasts, the SVMRBF method gave the most accurate results by a number of distinct metrics. Notably, the ensemble nowcast had lower error than each individual input stream. To my knowledge, this work is the first time that multiple digital surveillance streams have been used simultaneously to nowcast influenza. The resulting nowcasts appear to outperform most or all other previously published nowcasting systems. There are, however, serious limitations inherent to this methodology, including limited geographic resolution and susceptibility to missing data. I address these and other issues in Chapter 4.

Beyond nowcasting, there is widespread interest in *forecasting* disease outbreaks to minimize losses which would otherwise have been preventable given prior warning. In recent years, the CDC has sponsored two challenges (and a third currently underway) to predict influenza epidemics in the US [86, 87, 88], Defense Advanced Research Projects Agency (DARPA) has sponsored a challenge to predict the invasion of chikungunya into the Americas [89, 90], a consortium of government agencies in collaboration with the Office of Science and Technology Policy (OSTP) have sponsored a challenge to predict Dengue outbreaks in Puerto Rico and Peru [91], and the Research and Policy for Infectious Disease Dynamics (RAPIDD) group of the National Institutes of Health hosted a workshop for forecasting Ebola outbreaks [92]. As exemplified by the fields of meteorology and econometrics, statistical and computational models are frequently used to understand, describe, and forecast the evolution of complex dynamical systems [93, 94]. The situation in epidemiological forecasting is no different; data-driven forecasting frameworks have been developed in a variety of settings.

The epidemiological forecasting literature is large and rapidly growing. Several reviews have recently been published in an attempt to catalog and organize the vast amount of work being done in this relatively new field of research [95, 96, 97]. Here I focus specifically on one particular influenza forecasting system, the SIRS-EAKF framework [65]. To my knowledge it was the first to produce and publish seasonal influenza forecasts in real-time, and for 108 US cities. The model is an extension of the susceptible-infectious-recovered-susceptible (SIRS) model that is driven by absolute humidity (AH). More specifically, AH is used in the model to modulate the transmissibility of influenza, as there is empirical evidence that AH affects viral survival and transmission [33, 34]. This augmented SIRS model, functioning as a process model, is coupled with the Kalman Filter to estimate the system state—influenza incidence in a particular city. Ensembles of 200 SIRS-KF members were used in an Ensemble Adjustment Kalman Filter (EAKF) framework to track the mean and covariance of SIRS state variables and model parameters over time. In addition to retrospective forecasts, real-time season-wide forecasts were produced during the 2012–2013 flu season. To produce these forecasts, 150 such EAKF ensembles were

integrated through time up to the end of the flu season, providing a set of 150 trajectories of flu incidence within each city. The peak week of the epidemic was computed from each trajectory, and the mode of these peak weeks was used as a point prediction. Analysis of forecasting performance indicates that the SIRS-EAKF framework is far more accurate than methods based on resampling of historical trajectories. This methodology and initiative of publishing forecasts in real-time were significant improvements in the state of the art of epidemiological forecasting. However, the state of the art continues to improve, and I explore in detail several alternative approaches to epidemiological forecasting in Chapter 5.

One alternative approach to epidemiological forecasting is to build forecasts that are based on human judgment. In general, methods based on collective judgment take advantage of the interesting observation that group judgment is generally superior to individual judgment—a phenomenon commonly referred to as "The Wisdom of Crowds". This was prominently illustrated over a century ago when Francis Galton showed that a group of common people was able to guess the weight of a 1198 pound ox to within 9 pounds [98]. Since then, collective judgment has been used to predict outcomes in a number of diverse settings, including finance, economics, politics, sports, and meteorology [99, 100, 101]. A more specific type of collective judgment arises when the participants (whether human or otherwise) are domain experts—a "committee of experts". This approach is common in a variety of settings, for example in artificial intelligence and machine learning in the form of committee machines [102] and ensemble classifiers [103]. Other examples of human involvement in influenza research include prediction markets [75, 104] and participatory surveillance like Flu Near You [105, 106].

## 2.2 Innovation

Despite the significant advances in our understanding of the epidemiology and evolution of influenza, many questions remain unanswered [107]. It is unknown, for example, to what extent the generalized immunity theory accurately describes the true immune response against influenza. One of the biggest obstacles to addressing this concern is that generalized immunity is not well characterized, and in particular, it is generally assumed without empirical evidence to be powerful yet transient. I specifically address the latter of these issues in Chapter 3 by examining the role of generalized immunity in shaping both the evolutionary and epidemiological dynamics of influenza. In doing this, I more precisely define which regimes of human immunity are most plausible by assessing outcomes under various immunity assumptions with respect to a large set of targets for which there exists abundant empirical evidence.

In the US, we have the relative luxury of having significant historical and ongoing surveillance of influenza. However, this alone is not sufficient to answer the following critical and deceptively simple question: what is the current distribution and severity of flu in the US? Traditional surveillance, while invaluable for understanding the past, is unable to provide an answer to this question because of certain restrictions on geographic resolution and because of the delays inherent in clinical data collection and reporting. Looking forward, it is difficult to predict the course of an outbreak, even with complete situational awareness; it is even harder to make predictions when the current situation is not known to a reasonable level of accuracy. I address this critical shortcoming in our surveillance capability in Chapter 4 by constructing a novel indicator

of influenza activity in the US. Through a sensor fusion approach to the data assimilation problem, I show that a diverse set of digital surveillance proxies and a variety of prediction methods can be aggregated across space and time to provide accurate estimates of influenza within US states in real-time.

Given the magnitude of time, energy, and resources collectively invested by both participants and organizers in the numerous recent forecasting challenges, it is critical that qualitative and quantitative assessments be made to help understand where epidemiological forecasting excels and where it lags. To assess accuracy, forecasts are typically compared to predefined baselines and to other, often competing, forecasts. The focus has traditionally been on comparisons between data-driven methods. There has been less work toward understanding the utility of alternative approaches, including those based on human judgment. In Chapter 5 I develop and apply two distinct approaches to epidemiological forecasting, with special emphasis on a method based on collective human judgment. Each of these forecasting frameworks has been successfully put to the test though numerous forecasting challenges. I assess the accuracy of forecasts produced by these systems in several settings. Finally, I provide a demonstration of the state of the art by contrasting the performance of data-driven and human judgment methods for epidemiological forecasting, illustrating the relative advantages, and drawbacks, of each approach.

# Chapter 3

# Inferring Parameters of Human Immunity by Modeling Influenza

> Many key concepts concerning the nature of immunity have originated from the very practical need to control virus infections.
>
> *Peter Charles Doherty*

Much of this chapter is based on [108].

## 3.1 The evolutionary conundrum

The dynamics of rapidly evolving pathogens can be broadly divided into two classes: *epidemiological* and *evolutionary*—collectively referred to as "phylodynamics" [18]. Epidemiological dynamics include every aspect of disease that can be measured as a function of the host population: incidence, attack rate, reproductive number, and anything else that depends on the trajectory of an outbreak. Evolutionary dynamics on the other hand are the aspects of disease which are measured as a function of the pathogen population: rate of evolution, genetic and antigenic diversity, and anything else that can be derived from a phylogeny.

For decades, the epidemiological dynamics of flu have been well understood, evidenced by the fact that multiple generative models have been used to simulate these dynamics. This is perhaps unsurprising given that we have observed, for many decades, a regular and repeating pattern of winter epidemics in temperate regions. In fact, this pattern is used as a starting point for many forecasting strategies—a topic discussed in much more detail in Chapter 5. Even simple compartmental models provide a framework that is capable of generating, roughly, these outbreak patterns.

The story for influenza's evolutionary dynamics is, however, vastly different. Like other RNA viruses, the influenza viruses exhibit a rapid rate of genetic and antigenic drift. The A/H3N2 subtype in particular has the most rapid drift rate of the influenza viruses [109], and because of its volatile antigenic signature, it has been the most prevalent subtype in several recent flu seasons. With a rapid rate of evolution and a continual selective pressure to evade host immunity,

influenza could be reasonably expected to exhibit sustained, even saturating, viral diversity. Yet the opposite appears to be the case; since its pandemic appearance in 1968, A/H3N2 has given rise to a predominantly linear phylogeny, consisting of a single, well-defined trunk and short-lived side branches (Figure 3.1).

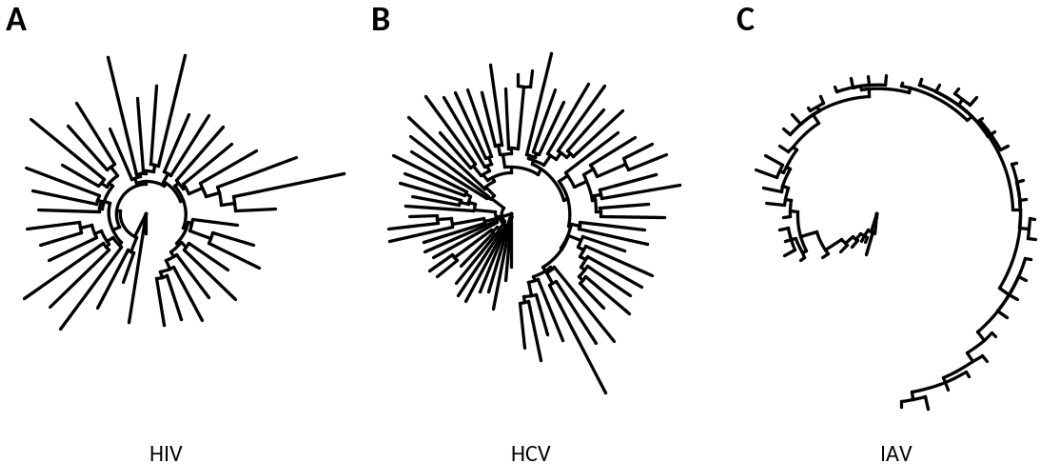


Figure 3.1: **Representative phylogenies of RNA viruses.** Subplots show maximum likelihood phylogenetic trees of selected RNA viruses, based on [18, 27]. **(A)** Human immunodeficiency virus (HIV); type 1. **(B)** Hepatitis C virus (HCV); type 1b. **(C)** Influenza A virus (IAV); type A/H3N2.

It was not until much later that the first generative models were posited that could, at least to some extent, explain the distinctive evolutionary dynamics of influenza. What may have been missing from the earlier models, it turns out, is a *hypothesized* key component of the immune response against influenza.

## 3.2 A working model of influenza

### 3.2.1 Description of the base model

One of the first models to successfully capture the combined phylodynamics of influenza leveraged a two-component immune response [42]. The first of these captures the classical notion of sustained protection against specific viral strains that a host has been previously exposed to. The second captures the more novel idea that a temporary protection against all strains is conferred immediately following exposure to any particular strain. This latter type of protection is known by several names, including *transient strain-transcenting immunity*, *temporary nonspecific immunity*, and *generalized immunity*.

The model described above (referred to going forward as the "base model") can be summarized as follows. It is an individual-based generative model of the long-term spread and evolution of influenza within a human population. Individuals are distributed according to a

semi-stochastic spatial hierarchy; they are assigned random coordinates drawn uniformly within a patch, with local clusters of hosts representing neighborhoods and the entire patch representing a large population center. Patches are arranged in a grid, and the top and bottom halves of the grid represent the northern and southern hemispheres, respectively. The likelihood of transmission between any pair of hosts is based on their proximity: within-neighborhood ($R_0 = 5$), within-patch ($R_0 = 0.4$), or cross-patch ($R_0 \approx 0.004$). A 25% sinusoidal seasonal forcing modulates transmissibility in opposite phase for the northern and southern hemispheres. Individual viral strains are defined by a small genome containing only the codons putatively under positive selection: 36 nucleotides encoding four hypothetical epitopes, each consisting of three amino acids. Antigenic phenotype is the set of 12 amino acids encoded by this genome, and under this parameterization there are 2012 possible distinct antigenic types. Following exposure to the virus, the probability of infection depends on the immune history of the exposed host. Upon infection, hosts incubate for two days, are infectious for four days, and are healthy afterward. When the host lifespan is reached, the host is respawned with a blank immune history. There is a small chance of mutation each day in each infected host ($p = 10^{-5}$ per nucleotide (NT)), with new mutants arising (replacing the parent strain in the infected host) and going extinct stochastically.

In the base model, immunity is conferred by two independent mechanisms: a long-term immunity against previously encountered strains and a short-term immunity against all strains. The long-term immunity provides very strong protection against strains that a host has been previously infected with and provides a weaker protection against strains that are similar to those known strains. In particular, the resistance afforded against a particular strain is a function of the number of novel amino acids encountered at each position, providing 99% resistance against strains with 2 or fewer novel amino acid substitutions (an immune escape threshold) and falling linearly to 25% resistance as the number of novel amino acid substitutions approaches saturation (12). The short-term immunity (generalized immunity) provides a broad protection against all strains and decays exponentially over time.

### 3.2.2 Questioning the mechanisms of immunity

Finally, a generative model was available that could account for many of the characteristic phylodynamics of influenza—but not all questions had been answered. In fact, this model raises an even bigger question: is this notion of a generalized immunity an actual part of the human immune response against influenza? On the surface it seems that to answer this question empirically would require a very strange experiment—exposing people to influenza and measuring the strength of their subsequent immune response. This is, of course, absurd. An alternative could be to test for such an immune response in an animal model, but this is also prohibitive for a number of the same reasons: measuring immune response is nontrivial, it is unclear how large of a sample would be needed, and the duration of the experiment would need to be at least on the order of the duration of protection of generalized immunity. Even under ideal conditions, it would require the very strong (and tenuous as best) assumption that the results of the animal system can be translated to a statement about the human system. At a very minimum, to even be able to perform such an experiment at all, we need a better and more specific hypothesis about the nature of generalized immunity.

There is, however, an alternative approach—one that could both give an idea of the plausibil-

ity of generalized immunity in humans and also more specifically define our notion of generalized immunity. Even though directly testing for generalized immunity in humans is currently infeasible, there are other measures for which we have abundant empirical data. These measures are of course the ones alluded to earlier: the measures based on the population-level epidemiological and evolutionary dynamics of influenza. The intuition for my approach is this: given a model that can produce outcomes on which the same measures can be computed, find the parameterizations of human immunity that give rise to the outcomes that are most congruent with what we expect to observe with influenza in reality. In other words, we know by a variety of metrics what the outcomes *should* look like; which model parameterizations, and in particular those of human immunity, produce these outcomes? This idea is similar in spirit to Approximate Bayesian Computation [110], whereby the likelihood of a given parameterization is inferred from the overall parsimony of its outcome.

### 3.2.3 Model implementation and extensions

I implemented an open-source simulator of the base model with the primary aim of studying how the parameterization of generalized immunity affects the outcomes predicted by the model [111, 112]. Taking advantage of a considerable increase in available computing power since the publication of the base model, I made several deviations from the default parameterizations to relax some of the strong assumptions that were previously required. Foremost among these assumptions were a small population size (12 million hosts) and a short host lifespan (30 years).

These parameter choices are less than ideal for several reasons, and naturally I would like to use values that reflect reality as closely as possible. One artifact of increasing the host lifespan in a small population is that the virus becomes prone to stochastic extinction. Another potential problem that arises with such small a population size is that in some situations adaptive immunity alone is able to constrain viral diversity [42, 113]. Although the original study primarily used the smaller population size, it was demonstrated (as a proof of concept) that reasonable outcomes could be obtained with a much larger population size (100 million hosts) and more reasonable lifespan.

The main obstacles facing an implementation with a population size of seven billion and a lifespan of around 60 years are computational in nature. There is an unavoidable trade-off between the number of hosts allocated and the amount of RAM and CPU time needed. Given the hardware available, and in the interest of time, I found that a population size of 100 million hosts was achievable for the bulk of my simulations, which numbered in total in the thousands. Since I ran all simulations at the larger population size, I was able to assume a more reasonable host lifespan of 60 years by reducing the volume of transmission between distinct geographical patches. These modest departures from the default parameterization of the base model result in a more reasonable approximation of reality, and I expect that these changes will result in a more reliable set of outcomes.

In fashion similar to the larger population proof of concept in [42], I was able to run a single instance with a population size of one billion hosts, suggesting that, given appropriate hardware (and an abundance of patience), the model could be used to simulate every living person.

## 3.3 Assessing outcomes

### 3.3.1 Epidemiological features

Consider the time series of an epidemic, consisting of incidence (the number, or proportion, of infected hosts) as a function of time. As previously discussed, there are many ways to describe such a trajectory, including features like the width, height, area, and overall shape of the epidemic. In what follows, I formalize each of these features and give the expected values for influenza A/H3N2 based on a survey of the relevant literature.

**Annual Attack Rate** (AAR) is the number of individuals infected during a single year, averaged over all years. This is effectively the cumulative incidence over a period of one year and is typically measured as the percentage of the population that becomes infected during the epidemic. Several estimates of the AAR of influenza are available in the literature [5, 29, 38, 39, 114, 115, 116] and give typical values ranging from $5\%$ to $25\%$, depending on age; I set the age-independent target value to $15\%(\pm 10\%)$ based on these estimates.

**Epidemic Duration** is a measure of how long seasonal epidemics last, averaged over all seasons. I defined this as the range of weeks, containing the week of peak incidence, that captures 90% of the seasonal attack rate. Based on my estimate of the empirical epidemic duration for the period spanning July 1, 2003 to July 1, 2012 using US surveillance data available from FluNet [117], and in accordance with similar estimates available in the literature [5, 24, 114, 115, 118], I set a target average epidemic duration of $12(\pm 2)$ weeks.

**Reproductive Number** ($R_p$) is a dimensionless number that quantifies the expected number of secondary cases arising from a primary case throughout the duration of an infectious period. This quantity differs from the basic reproductive number ($R_0$) in that $R_0$ assumes a completely naive population, whereas $R_p$ assumes a population with some pre-existing partial immunity [38]. In the case of influenza, where a significant portion of the population has lingering immunity from strains encountered either during a previous season or through vaccination, $R_p$ is a more appropriate measure than $R_0$. $R_p$ for seasonal influenza in temperate regions has been estimated to be $1.3$ (95% CI $1.2$–$1.4$) [28, 38], though here I assumed the more permissive range of $1.1$–$1.5$ to be credible.

**Peak Weekly Incidence** is the peak of the weekly incidence of each seasonal epidemic, averaged over all seasons. The true peak incidence of influenza is difficult to measure, so instead I base the target range on estimates from the previously mentioned individual based models. Specifically, I estimate based on the model described in [42] the peak weekly incidence under normal epidemic conditions to be $2,500(\pm 1,500)$ hosts per $100,000$ hosts , or $2.5\%(\pm 1.5\%)$.

### 3.3.2 Evolutionary features

Considering the evolutionary history of the virus over time, there are several ways to describe the course of viral evolution. These features capture the ideas of diversity, rate of mutation, and various aspects of viral lineage. As before, I give the expected values of these features with respect to the empirically observed lineage of influenza A/H3N2.

**Pairwise Diversity** is the prevalence-weighted mean pairwise number of amino acid differences between all pairs of strains existing on some day, averaged over all days. It has been previously used to quantify the average amount of viral antigenic [42] and genetic [47] diversity and is an indicator of whether viral diversity is constrained (consistent with influenza) or unconstrained (inconsistent with influenza). Because it is the average number of amino acid differences between two strains, the possible values range from $0$ to the number of amino acids modeled, $12$ in these simulations. Though the diversity of influenza follows a "boom-and-bust" pattern [47], I expect that pairwise diversity should, on average, remain low. Therefore, I assumed the plausible target value for these modeled strains to be $2(\pm 1)$ amino acids.

**Fixation Rate** is a measure of how quickly the virus evolves, as indicated by the number of novel mutations becoming fixed in the viral phylogeny over some period of time. The mutation rate for the A/H3N2 subtype is particularly high [119] and, for the sites assumed to be under positive selection, has been measured to be $0.053(\pm 0.01)$ substitutions per site per year [42].

**Most Recent Common Ancestor** (MRCA) measures the number of years separating all contemporaneous strains. Here I used the average number of years of evolution separating two randomly sampled strains on any given day to approximate the MRCA, as in [25, 120]. The target value for influenza, using the HA gene only, is roughly $1.12$ years with a 95% confidence interval roughly spanning $0.58$–$1.97$ years (excluding the 2002-03 season outlier) [24].

**Kappa** ($\kappa$) is a dimensionless number which quantifies the potential for antigenic evolution of rapidly evolving viruses [27]. I use $\kappa$ to probabilistically quantify where on the diversification spectrum, from very constrained (as expected for A/H3N2 flu) to exceedingly diverse (as observed for between-host HIV), a simulated phylogeny falls. $\kappa$ is measured here as the parameter of the best fit Poisson distribution (determined by maximum likelihood estimation) given the counts of excess antigenic variants, per variant, over the duration of the simulation. With its characteristic phylogeny, influenza exhibits a relatively low degree of phylogenetic branching, and $\kappa$ has been estimated to be $0.11$ with a 95% confidence interval roughly spanning from $0.01$–$0.49$ based on [27].

## 3.4 Mapping the parameter space of generalized immunity

In the model, generalized immunity is controlled by two parameters: *strength* and *duration*. The strength of generalized immunity ($\omega$) is a dimensionless number ranging from 0 to 1, which directly translates to the maximum initial probability of immunity against *all* viral phenotypes. The duration of generalized immunity ($\tau$) is the half-life (in units of time) controlling the rate at which the overall protection of generalized immunity decays. After some amount of time has elapsed ($\Delta t$), the probability of protection due to generalized immunity is:

$$\Pr_{\text{GI}}(\Delta t) = \omega e^{\left\{-\Delta t/\tau\right\}}$$

Together with the protection of cross immunity ($\Pr_{\text{CI}}$), a function of antigenic distance ($f(D)$),

the probability of *infection* is given by:

$$\Pr_{\text{infection}}(\Delta\tau, f(D)) = \left(1 - \Pr_{\text{GI}}(\Delta t)\right)\left(1 - \Pr_{\text{CI}}(f(D))\right)$$

Since the empirical existence of a generalized immune response is not entirely certain, I wanted to find which parameterizations could at least be plausible. To do this, I set out to map the two-dimensional parameter space of generalized immunity. The space in its entirety, however, contains several uninteresting regions. For practical considerations, I constrained the exploration to the region bounded by strength from $0.25$ to $1.00$ and half-life from $41$ days to $60$ years, reasoning that **a)** exceedingly weak or short parameterizations approach the degenerate case of nonexistence of generalized immunity and that **b)** parameterizations lasting longer than the lifespan of a host effectively confer complete and permanent immunity against all strains. Over this bounded space I imposed a grid of sixty points, roughly equally spaced in strength and in the logarithm of half-life, to represent a large set of potential parameter regimes of generalized immunity.

Using the simulator I implemented, I "sampled" each grid point (parameterization) twenty times. Each individual sample was a fifty year simulation of influenza transmission and evolution in a population of 100 million human hosts. From the output of each simulation, I measured each of the previously discussed epidemiological and evolutionary outcomes, which are summarized in Table 3.1.

| Measure | Target Value | 95% CI | Source |
|---|---|---|---|
| Annual Attack Rate (AAR) | 15% | 5–25% | [5, 29, 38, 39, 114, 115, 116] |
| Epidemic Duration | 12 weeks | 10–14 weeks | [5, 24, 114, 115, 118] |
| Reproductive Number ($R_p$) | 1.3 | 1.1–1.5 | [28, 38] |
| Peak Weekly Incidence | 2.5% | 1–4% | following [42] |
| Pairwise Diversity | 2 AA | 1–3 AA | following [42] see also [47] |
| Fixation Rate | 0.053 NT/yr | 0.043–0.063 NT/yr | [42] |
| Most Recent Common Ancestor (MRCA) | 1.12 yr | 0.58–1.97 yr | [24] see also [25, 120] |
| Kappa ($\kappa$) | 0.11 | 0.01–0.49 | based on [27] |

Table 3.1: **Summary of epidemiological and evolutionary measures.**

Examining parameter space maps with respect to each epidemiological (Figure 3.2) and evolutionary (Figure 3.3) outcome brings to light several interesting trends. To point out some of these trends, it is helpful to first consider the behavior of the model within the regimes of extreme

parameterizations: strengths that lead to either ineffective or perfect protection and durations that lead to either ephemeral or permanent protection.

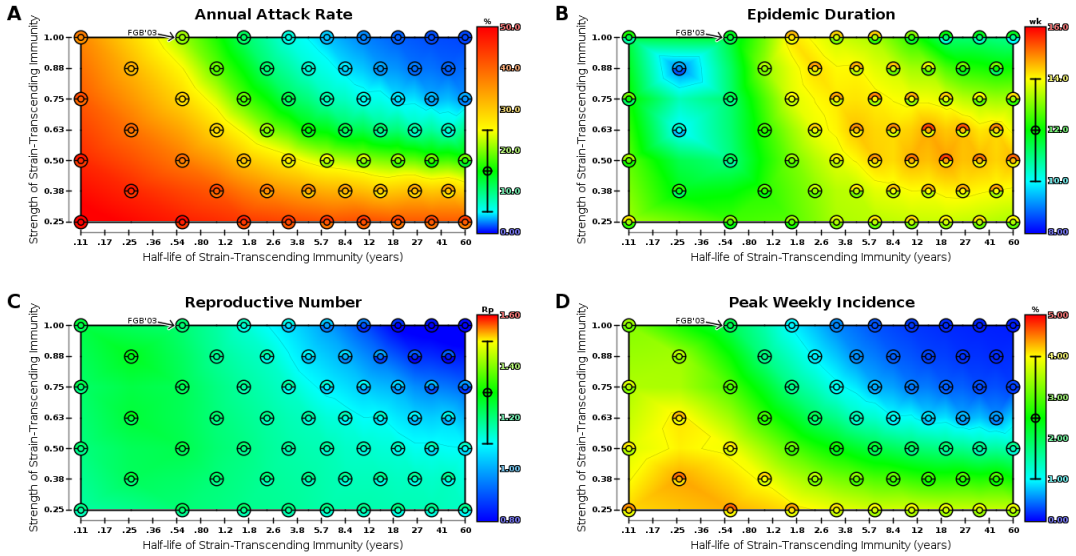


Figure 3.2: **Epidemiological parsimony is measured across generalized immunity parameter space.** Sixty parameterizations of generalized immunity (strength and duration) were simulated. Within each map, the sixty parameterizations are represented by a set of circles and semicircles; the inner circle at each point represents the sample mean of the measure, and the top and bottom semi-circles represent the mean plus and minus one sample standard deviation, respectively ($n = 20$ realizations for each point). Color corresponds to the agreement between the simulated outcome and the expected outcome for influenza; blue indicates a value below the 95% CI for influenza, red indicates a value above the 95% CI for influenza, and green represents the expected value for influenza. The 95% CI for influenza is marked on each color scale, and the target value is indicated by "$\oplus$". The tip of the "FGB'03" arrow indicates the default parameterization given in [42]. Triangulation and interpolation were used to achieve smooth shading throughout the space to facilitate visual identification of spatial trends. Faint contour lines demarcate confidence interval boundaries. **(A)** AAR measured from model output; target value is $0.15$ (95% CI: $\pm 0.10$). **(B)** Epidemic duration measured from model output; target value is $12$ (95% CI: $\pm 2$) weeks. **(C)** $R_p$ measured from model output; target value is $1.3$ (95% CI: $\pm 0.2$). **(D)** Peak weekly incidence measured from model output; target value is $0.025$ (95% CI: $\pm 0.015$).

Consider first the regime most similar to a model without generalized immunity, where protection is both ineffective and ephemeral (bottom-left in maps). Without the extra protection afforded by generalized immunity, hosts are much more susceptible to repeated infection. This is reflected by a high incidence and a very high attack rate, although epidemic duration and reproductive number are within their target ranges. As a result of an increase in the number

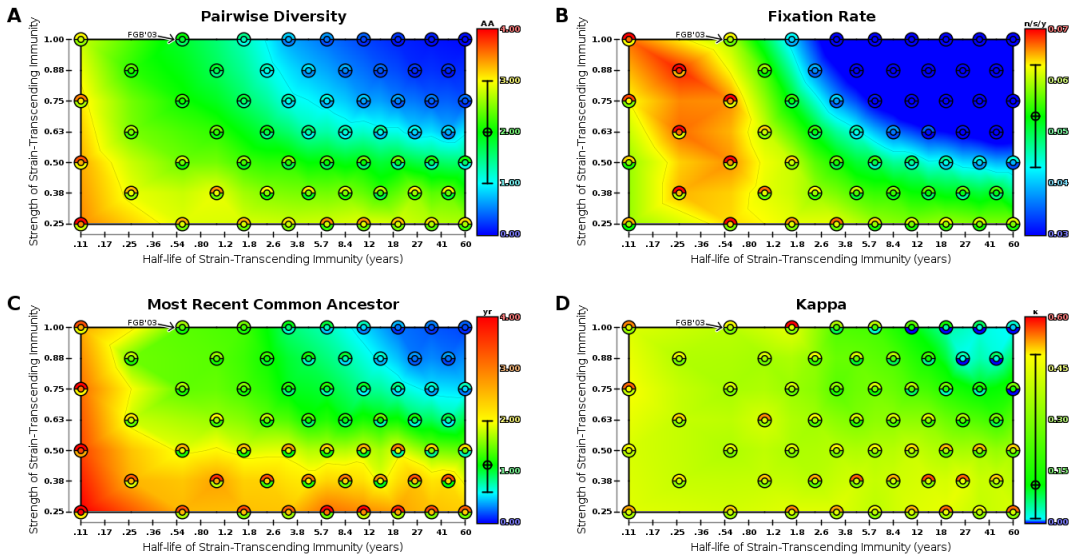


Figure 3.3: **Evolutionary parsimony is measured across generalized immunity parameter space.** Maps are as described in Figure 3.2. **(A)** Pairwise diversity measured from model output; target value is 2 (95% CI: $\pm 1$) amino acids. **(B)** Fixation rate measured from model output; target value is $0.053$ (95% CI: $\pm 0.01$) substitutions per site per year. **(C)** MRCA measured from model output; target value is $1.12$ (95% CI: $0.58$–$1.97$) years. **(D)** $\kappa$ measured from model output; target value is $0.11$ (95% CI: $0.01$–$0.49$).

of infections, the virus has more opportunity to diversify. Furthermore, pairwise diversity and MRCA are particularly high, while $\kappa$ and fixation rate are elevated, though still within their target ranges. Selective pressure within this regime is weak, as minor antigenic changes are sufficient to evade the similarity-based protection of cross-immunity. A characteristic trajectory within this parameter regime is shown in (Figure 3.4 A); peak incidence and attack rate are greatly elevated, diversity is unconstrained, and phylogeny is exceedingly branched.

Perhaps the most extreme regime is that of perfect and permanent generalized immunity (top-right in maps). Here there seems to be an antipodal response, both epidemiologically and evolutionarily. With such extreme protection, hosts are essentially protected for life following an initial exposure to *any* viral phenotype. As can be expected, extreme lows are reported for almost all measures. Epidemic duration, which is below its target but within its plausible range, is the sole exception, presumably because the width (though not the magnitude) of the epidemic curve remains on average unchanged. Due to a dearth of susceptible hosts, the virus is saved from stochastic extinction only by the eventual accumulation of newborn hosts in a manner reminiscent of measles. A characteristic trajectory within this parameter regime is shown in (Figure 3.4 B); sporadic outbreaks result in greatly reduced peak incidence and attack rate, pairwise diversity approaches zero during extended periods of near-extinction, and phylogeny is exceedingly slender.

Next, consider the two hybrid regimes of generalized immunity: ineffective but permanent, and perfect but ephemeral (bottom-right and top-left in maps, respectively). For the first time, reasonable outcomes are observed for many measures in these regimes, and in particular both incidence and diversity are satisfied within and between both regimes. According to most maps, the value being measured is relatively static throughout the transition from corner to corner. One exception to this trend is epidemic duration, which generally transitions from low (short) to high (long) as the duration of generalized immunity increases. Another exception is fixation rate, which indicates very strong positive selection in the perfect-but-ephemeral regime and influenza-like positive selection in the ineffective-but-permanent regime. The latter result indicates that very short durations of generalized immunity pressure the virus into a state of rapid, stepwise change in antigenicity, whereas very long durations of generalized immunity pressure the virus into making less rapid, but more significant, jumps in antigenicity. As the strength of generalized immunity approaches its minimum, selective pressure depends only on adaptive immunity, which is, at least in the model, a nonlinear function of antigenic distance. As a result of this nonlinearity, rare-but-large antigenic steps are more favorable than frequent-but-small antigenic steps.

Finally, I come to what could be the plausible regime of generalized immunity: moderate in strength and of intermediate duration (having a half-life on the order of several months to years). All eight measures are generally close to their target values when the model is parameterized within this regime. Characteristic trajectories within this parameter regime are shown in Figure 3.4 C and Figure 3.4 D; regular annual epidemics with an attack rate of around 15% are observed, pairwise diversity takes on the familiar pattern of gradual build-up followed by rapid collapse coinciding with extinction of prominent lineages, and phylogeny is generally linear over long time spans, with a moderate amount of short-term branching.

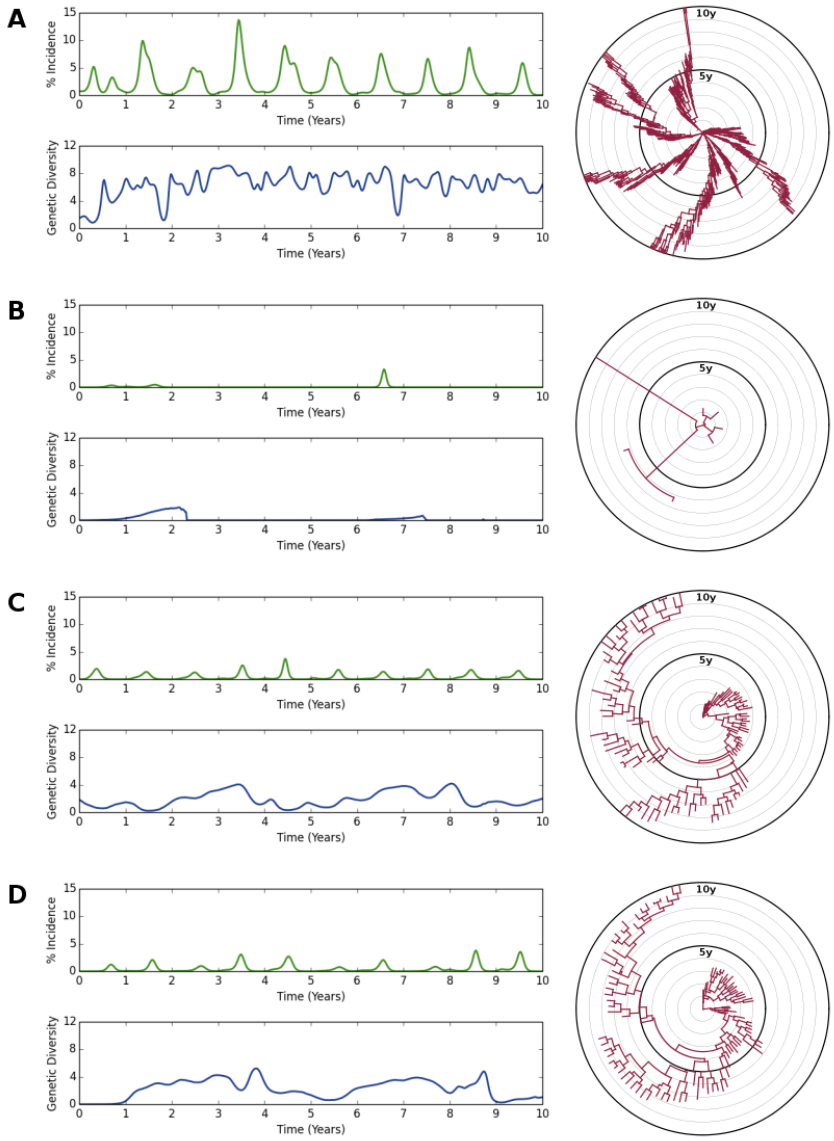


Figure 3.4: **Characteristic epidemiological and evolutionary dynamics are observed within different parameter regimes.** For each parameterization, a ten year excerpt from a typical simulation is shown. Graphs on the left show weekly percent incidence (northern hemisphere only, green) and weekly mean pairwise antigenic diversity (blue). Radial dendrograms on the right show phylogeny (red) with concentric rings marking one year intervals. **(A)** Weak and short-lived generalized immunity (strength $= 25\%$, half-life $\approx 1.3$ months). **(B)** Strong and long-lived generalized immunity (strength $= 100\%$, half-life $\approx 11$ years). **(C)** Parameterization similar to that described in [42]: strong and short-lived generalized immunity (strength $= 100\%$, half-life $\approx 7$ months). **(D)** Alternative parameterization within the plausible region: moderate strength and intermediate duration generalized immunity (strength $\approx 63\%$, half-life $\approx 28$ months).

Overall, I observe that:

- AAR is closest to the target value when strength and duration of generalized immunity are both relatively high, with some degree of trade-off (Figure 3.2 A).
- Average epidemic duration is a complex function of generalized immunity, but in general is close to the target value for moderate durations of generalized immunity (Figure 3.2 B).
- $R_p$ is closest to the target value when generalized immunity is strong and short-lived (Figure 3.2 C).
- Peak weekly incidence is closest to the target value when strength and duration of generalized immunity are inversely proportional (Figure 3.2 D).
- Pairwise diversity is closest to the target value when strength and duration of generalized immunity are inversely proportional (Figure 3.3 A).
- Fixation rate is closest to the target value when strength and duration of generalized immunity are inversely proportional (Figure 3.3 B).
- MRCA is closest to the target value when strength of generalized immunity is moderate or high; this measure appears to be somewhat insensitive to changes in duration of generalized immunity (Figure 3.3 C).
- $\kappa$ is closest to the target value when generalized immunity is relatively strong over a moderate duration (Figure 3.3 D).

## 3.5 Computing likelihood across parameter space

Although it is informative to examine each measure in isolation, the goal now is to characterize the plausible range of parameterizations of generalized immunity based on the cumulative evidence provided by all eight measures. In other words, for each point in parameter space, I want to use the evidence from all measures to form a combined estimate of the joint likelihood of each parameterization. I have a point estimate and a 95% credible interval for the empirical value of each measure, and I want to produce a single figure of merit. If I assume that all target distributions are Gaussian, then the Mahalanobis distance [121] can be used to calculate the joint likelihood I seek. The Mahalanobis distance measures the covariance-adjusted distance from a given point ($x \in \mathbb{R}^p$) to the center of a multivariate Gaussian distribution ($\mathcal{N}(\mu, \Sigma); \mu \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p \times p}$) as:

$$\text{MD}(x; \mu, \Sigma) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$$

Here $p$ is the number of features, $x$ is the vector of values measured from simulator output and $\mu$ is the vector of empirical target values for each feature. If I further assume that all features are independent, then the covariance matrix $\Sigma$ is the matrix whose diagonal entries are based on the width of each measure's credible interval and zero elsewhere. Given a Mahalanobis distance, the likelihood (and therefore the statistical significance) of a parameterization can be determined through the property that squared Mahalanobis distance is chi-squared distributed with $p$ degrees of freedom. After applying this methodology across parameter space, I arrive at the final map of plausible parameterizations of generalized immunity (Figure 3.5).

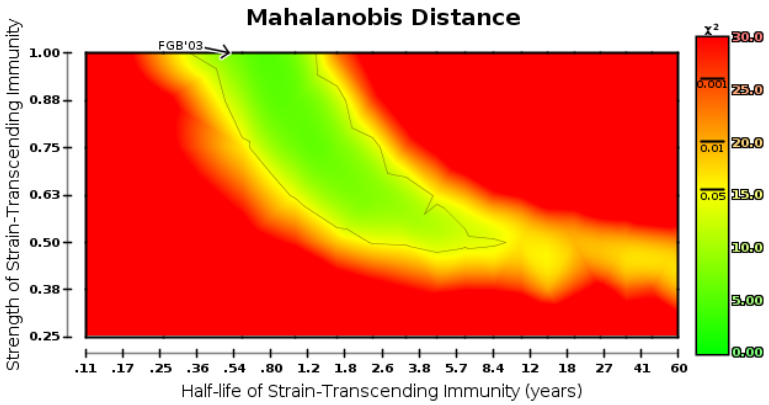


Figure 3.5: **Generalized immunity is potentially longer lasting and weaker than previously suspected.** Map is as described in Figure 3.2 with circles at each individual parameterization suppressed. Color encodes the value of the chi-squared test statistic; green represents inability to reject the null hypothesis that an outcome is not a multivariate outlier (equivalently, is influenza-like), and red represents high confidence in the rejection of that hypothesis. Lines on the color scale indicate probability cutoffs at the $p \leq 0.05$, $p \leq 0.01$, and $p \leq 0.001$ levels for a chi-squared distribution with eight degrees of freedom.

The original parameterization of generalized immunity given in [42] ("FGB'03") is very near to—but within—the threshold of significance. The bounded area of parameter space can be loosely interpreted as the set of parameterizations which are at least as likely as the original. The FGB'03 parameterization is just one extreme of a spectrum of plausible parameterizations of generalized immunity, which extends much further in both dimensions, up to a half-life of many years at half strength.

This probably raises a question: according to the model, what is the single most likely parameterization of generalized immunity? This is difficult to answer for several reasons. First, parameter space is continuous, but it was sampled over a discrete grid. Second, each sample contains some amount of noise due to the stochastic nature of each simulation, which makes it difficult to say precisely which points have the maximum likelihood. Third, there is no guarantee that the likelihood function is convex over parameter space, which means that there is no efficient method for finding the global optimum. Finally, and most importantly, the shape of the plausible region depends on every other parameter in the model and on which features are measured on model output—this will be explored in the following section.

With these caveats in mind, it is possible to give a rough estimate of the most likely parameterization. The grid can be made continuous through interpolation, each point was sampled 20 times to reduce the variance of the estimated likelihood, a brute force solution can be used to search for the point of maximum likelihood, and likelihood is reasonably robust to model parameterizations and output features. In Figure 3.5, the grid point with maximum likelihood (equivalently, minimum Mahalanobis distance) is at $\tau = 1$ year, $\omega = 87.5\%$.

## 3.6 Sensitivity analysis and robustness

### 3.6.1 Relaxing assumptions

When computing likelihood in the previous section, I made two simplifying assumptions: that the credible intervals of empirical targets are normally distributed and that all features are marginally independent. These assumptions are, of course, quite unrealistic. $\kappa$, for example, has a very skewed distribution, and the features of peak weekly incidence and annual attack rate, for example, are almost certainly correlated to some extent. Unfortunately, I do not have an analytical expression for the distribution of any of the features, and there is not enough empirical evidence available to reliably estimate the correlation structure between all features. Instead, I attempt to relax these assumptions by excluding subsets of features and estimating correlations from simulated outcomes, and I show how these changes affect my conclusions on the plausibility of generalized immunity.

To begin, note that most features have *effectively* normally distributed credible intervals. In fact, the only features which are explicitly not normally distributed are MRCA and $\kappa$. I now ask the following question: what are the plausible parameterizations of generalized immunity in the *absence* of MRCA and $\kappa$? To do this, I repeat the process of determining likelihood across parameter space and assess likelihood in terms of a chi-squared distribution with now only six degrees of freedom. There appear to be only minor changes to the region of plausible parameterizations, suggesting again that the intermediate regime of generalized immunity is most plausible (Figure 3.6 A).

Next, consider the assumption of independence among features. Here I take two separate approaches in an attempt to understand how this assumption affects my conclusions. First, I exclude measures which are suspected to be highly correlated. Second, I estimate the covariance matrix of the features to calculate a more stringent Mahalanobis distance.

Regarding the first approach, I exclude two features which are presumably highly correlated with some subset of the remaining features: peak weekly incidence and pairwise diversity. Of the peak weekly incidence (roughly measuring height), epidemic duration (rougly measuring width), and annual attack rate (roughly measuring area) triplet, any individual measure can be reasonably estimated given the other two, assuming that there is some archetypal epidemic shape—an idea applied in Chapter 4. Of these three related measures, I exclude peak weekly incidence because I am more confident in the target distributions of epidemic duration and annual attack rate. Pairwise antigenic diversity and MRCA both essentially measure the amount of divergence among extant strains. Given a distant MRCA, high diversity is not surprising; given a recent MRCA, low diversity is not surprising; and *vice versa*. I exclude pairwise diversity because true MRCA can be reasonably estimated from empirical sequence data, whereas it is unclear how to estimate, or to even define, pairwise diversity using the same data. As before, I evaluate likelihood across parameter space and observe only minor variations on the shape of the plausible region (Figure 3.6 B).

Regarding the second approach, I first estimate the sample correlation among features on simulated data. I generate this estimate separately at each grid point (recall that $n = 20$ simulations at each point) and observe that much of the correlation structure is preserved across parameter space. I illustrate this correlation structure below, showing the average of all individual correla-

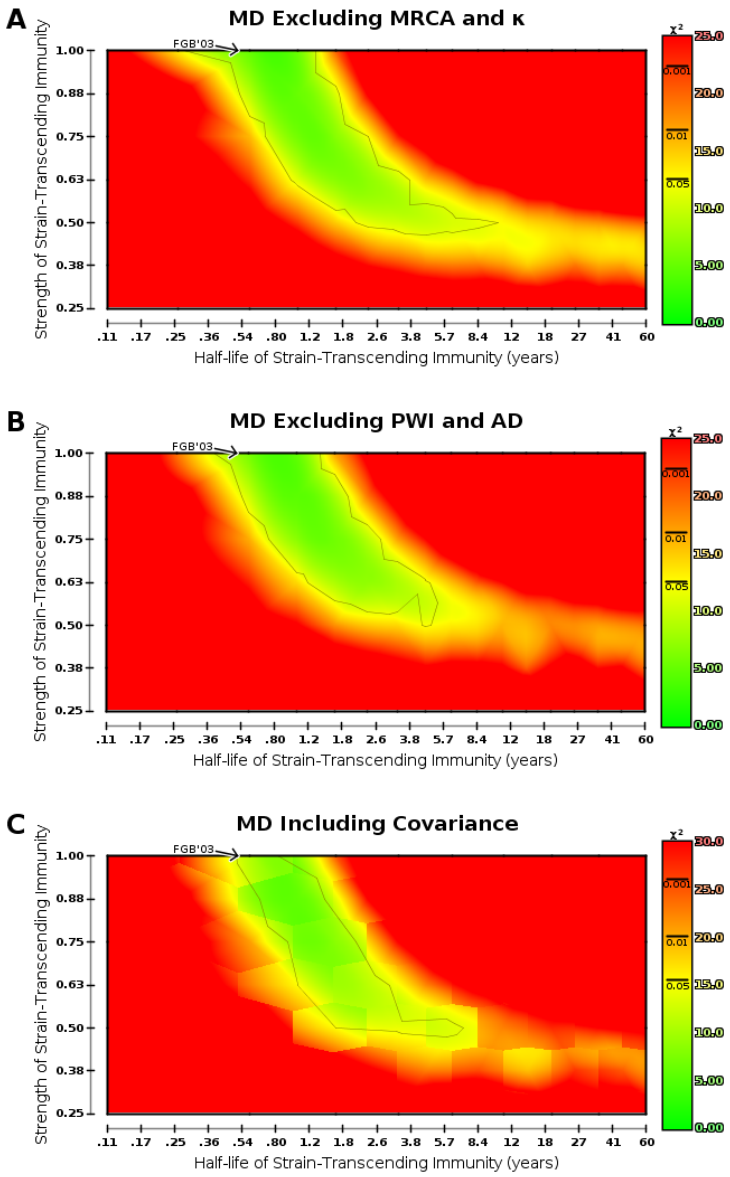


Figure 3.6: **Conclusions regarding generalized immunity are generally unchanged by relaxing assumptions.** Maps are as described in Figure 3.5 with threshold of significance adjusted appropriately. Maps show likelihood where **(A)** features with credible intervals not normally distributed are excluded, **(B)** features with the strongest correlations are excluded, and **(C)** Mahalanobis distance is computed using a feature covariance matrix estimated from simulator output.

tion matrices (Figure 3.7). Until now I implicitly employed a diagonal covariance matrix when calculating Mahalanobis distance, where diagonal elements were variances and non-diagonal elements were zero (no correlation between features). Now, I build a more informative covariance

matrix by using simulator output to estimate correlations between features. I generate a covariance matrix separately at each grid point by first converting the estimated correlation matrix into a covariance matrix and then element-wise averaging the old (diagonal) and new (non-diagonal) covariance matrices. Note that the diagonals of each matrix are equal, but the non-diagonals take 50% of the correlation structure to avoid overfitting. I use the resulting covariance matrices to calculate a Mahalanobis distance which takes into account correlations between all measures. Again, the shape of the plausible region remains overall unchanged (Figure 3.6 C).

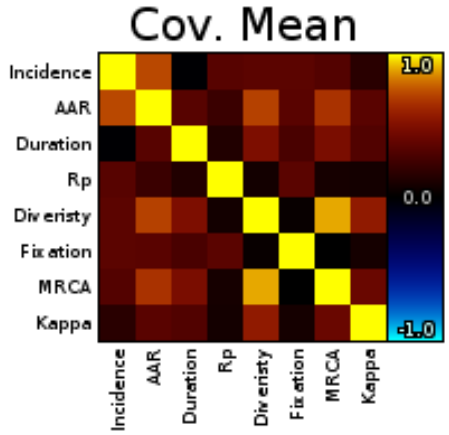


Figure 3.7: **Average correlation structure of features measured on simulator output.** The strongest correlations are on average between MRCA and pairwise diversity: 0.90; annual attack rate ("AAR") and PWI ("incidence"): 0.73; AAR and pairwise diversity: 0.71; and AAR and MRCA: 0.67.

Finally, it is interesting to take the intersection of the plausible region between each of the maps built with a different set of assumptions. In all, the original parameterization proposed by [42] is within—but very near to the border of—the plausible region. Additionally, all maps show the same general trend, and the shape of the plausible region is generally well conserved. The most extreme difference is probably the clipping of the plausible duration of generalized immunity. Whereas the original (Figure 3.5) map suggested that half-life parameterizations of up to roughly 10 years were plausible, these more stringent maps (particularly Figure 3.5 B) seem to truncate the extremes of the plausible region to a half-life of roughly 5–6 years. While this may seem at first like a significant difference, the original claim still holds: generalized immunity could plausibly be much weaker and longer lasting than commonly assumed.

### 3.6.2 Targeting median age instead of life expectancy

Global life expectancy is on the order of 60–70 years [122], and this is the value that I matched in my simulations. However, because the simulated population is unable to accommodate a growing population, it is also interesting to simulate a lifespan that is closer to the global median age, which is on the order of 30 years [123]. To determine to what extent shorter host lifespans affect my results regarding the plausibility of generalized immunity, I remapped a large portion

of parameter space under a host lifespan of 30 years. I ran sets of at least 5 replicates across the most salient portion of parameter space to determine how the shape of the plausible region is affected. As expected, reducing lifespan (equivalently, increasing host respawn frequency) results in an overall increase in the number of naive hosts, and attack rates are uniformly, but not unreasonably, increased compared to those of the simulations using 60 year lifespans (Figure 3.8 A). Although the shape of the plausible region changes slightly, I find that the primary conclusions still hold (Figure 3.8 B).

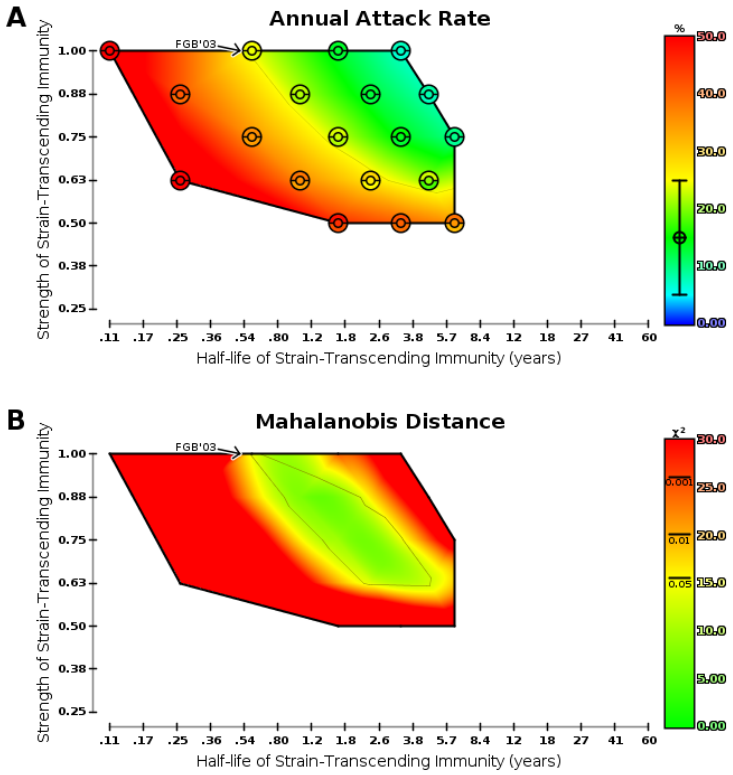


Figure 3.8: **Conclusions regarding generalized immunity are generally unchanged by using median age instead of expected lifespan.** Simulations used a lifespan of 30 years instead of 60 years ($n \geq 5$). **(A)** Attack rate, as in Figure 3.2 A. **(B)** Likelihood across parameter space, as in Figure 3.5.

### 3.6.3 A more realistic population structure

The simple world of the model consists of two equally populated hemispheres which are subjected to sinusoidal seasonal forcing in opposite phase. In the real world, seasonal forcing is much more complex, and only the temperate regions of the world experience the type of seasonal forcing that can be described by a simple sinusoid with an annual peak in the winter months. The highly-populated tropical regions of the world are not subject to such well-defined forcing, yet

it has been shown that strains arising and circulating in these tropical regions play a significant role in the diversification and spread of flu globally [25, 29]. It is therefore of great interest what impact the simple geography of the model has on immunity results and how a more realistic model of the world would change those results.

To this end I made a straightforward extension to the model to allow for patches representing tropical regions. These patches differ from the temperate patches only in that there is no imposed seasonal forcing. Additionally I redistributed the 20 patches (originally divided into 10 northern and 10 southern patches) into 8 northern, 10 tropical, and 2 southern patches to more closely approximate the actual distribution of the world population. All other aspects of the model, including between-patch contact rates, remained the same. As with the sensitivity analysis of host lifespan, I ran sets of 5 replicates across a broad range of parameter space to assess the impact of including tropical dynamics on the shape of the plausible region of immune parameters. I find that the shape of the plausible region changes somewhat more significantly, in particular becoming more jagged (presumably an artifact of the course grid resolution and a small number of replicates), and more importantly only extending up to a half-life of 2 years (Figure 3.9). Despite these moderate changes to the shape of the plausible region under a remarkably different model of the world, the conclusion remains in essence the same: generalized immunity could plausibly persist with a half-life that is meaningfully longer than originally anticipated.

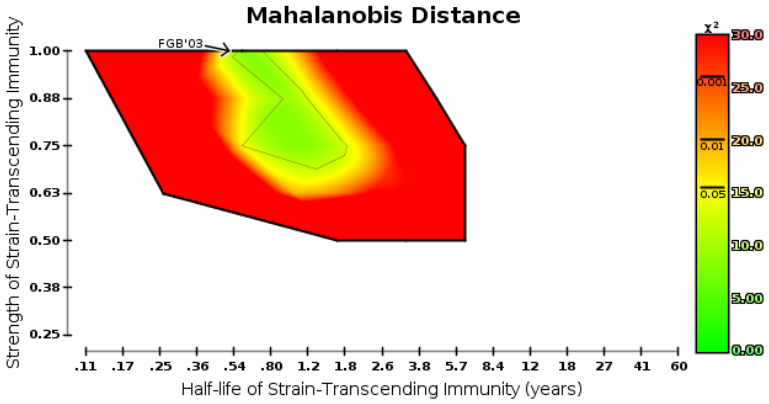


Figure 3.9: **Conclusions regarding generalized immunity are generally unchanged when using a more realistic population structure.** Simulations included a large tropical deme with no seasonal forcing ($n \geq 5$). Likelihood across parameter is shown space as in Figure 3.5.

## 3.7 Final considerations

Although it has yet to be conclusively shown that generalized immunity is responsible for influenza's characteristic phylodynamics, there have been several recent studies providing empirical evidence that cellular immune responses are able to confer some degree of heterosubtypic immunity. In mice, cellular immunity has been shown to reduce illness [124] and protect against

lethal infection following infection with different subtypes [125, 126]. Similarly, cell-mediated heterologous immunity has been observed in ferrets [127, 128]. Further, it has been suggested that weak heterologous immunity in humans is necessary to account for suppressed influenza B outbreaks following severe epidemics of influenza A [129].

Regarding the nature of antigenic evolution in the model, it is important to note that there are implicit constraints on the magnitude of antigenic change between viral progenitor and progeny strains which determine the process by which immune escape mutants arise. Antigenic evolution is concomitant with non-synonymous substitutions in viral codons, and although individual mutations can cause at most one amino acid substitution, mutations events are independent across nucleotides and can potentially occur simultaneously within an infected host on each day of infection. Although technically possible, it is exceedingly unlikely that any nascent mutant will differ from its parent strain at more than two amino acid sites, and because at least three amino acid substitutions are required to cross the immune escape threshold (in the current parameterization), any large antigenic changes will almost surely be the result of a series of mutation events over multiple days and across multiple hosts. While it has been demonstrated that models allowing for rare, but abrupt, changes in antigenicity are able to reproduce many of influenza's characteristic dynamics without a need for generalized immunity, the incremental changes in antigenicity modeled here are more consistent with our understanding of the empirical mechanisms of antigenic drift on the micro scale of short-term, within-host evolution [130]. However, on longer timescales it is generally understood that punctuated changes in antigenicity (which the model explored in this chapter neither prescribes nor proscribes) drive the cluster transitions observed roughly every 2-5 years with influenza A/H3N2 [19, 47, 131].

Although the model successfully recapitulates the phylodynamics of influenza A/H3N2, it should be noted that there are certain design choices which complicate the interpretation of simulated trajectories. The current model assumes a universe consisting of only influenza A/H3N2, and is therefore unable to generate the seasonal variations in dominant subtypes observed empirically. Complicating the addition of these additional strain types is our limited understanding of the strengths and durations of the interactions between them. Another shortcoming of the present model is an absence of the effects of broad vaccination campaigns routinely used in many parts of the world. Mass vaccination against contemporary strains undoubtedly has a non-trivial effect on the shape of the epidemic trajectory which cannot be captured in the simplistic universe of the model.

Additional improvements to the model could include using a more realistic population structure, perhaps taken from synthetic population estimates [132]; and incorporating climatological data, such as absolute humidity [34], to more accurately modulate transmissibility than the current sinusoidal forcing function. Finally, the simplistic, two-component immune system modeled here is a useful, but limited, abstraction of the complex and not entirely understood human immune response against rapidly evolving, antigenically variable viruses. For example, the confounding phenomena of original antigenic sin, antigenic seniority, antigen trapping, and back-boosting have a significant and non-trivial role in determining immune response [133, 134], yet none of these effects are explicitly modeled. (Although the effects of transient generalized immunity are arguably similar in some respects to the effects of back-boosting, which increases antibody titers against all previously encountered strains for a duration of time on the order of one year.) However, it would be straightforward to extend the model to emulate original anti-

genic sin and antigen trapping, and modeling the effect of back-boosting in a partially vaccinated population would be another interesting, albeit more challenging, direction for future work.

# Chapter 4

# Nowcasting Influenza through Sensor Fusion of Digital Surveillance

> Whenever the state of a system must be estimated from noisy sensor information, some kind of state estimator is employed to fuse the data from different sensors together to produce an accurate estimate of the true system state.
>
> *Simon J. Julier*
> *Jeffrey K. Uhlmann*

Much of this chapter is based on [135].

## 4.1 Situational awareness for preparedness

Epidemiologists, clinicians, and public health policy makers together face a serious and unique challenge; at any given point in time, the true extent of a disease outbreak is not fully known—perhaps even unknowable. This is in stark contrast with meteorologists, for example, who have at their immediate disposal a large set of highly accurate and real-time tools for data acquisition, including radar, atmospheric and oceanic sensor arrays, and satellite imaging. A similar parallel exists to some extent for economists who have in real-time a precise set of market indicators from a variety of reliable sources. It is desirable, advantageous, and profitable to have accurate and timely situational awareness in order to prepare for, and to mitigate the effects of, disease outbreaks, weather systems, and market fluctuations. There is, however, a fundamental difference between these scenarios, and it has to do with what can be learned in real-time from the state of the system under study.

Epidemiology is concerned with the spatial and temporal spread of disease through a host population. This is unfortunately difficult to measure, especially instantaneously. For flu in the US, matters are further complicated by the fact that a large fraction of infections are asymptomatic, those which present with symptoms often do not seek treatment, only the most severe

cases result in hospitalization, reporting of cases is in large part voluntary, the sample of reported cases is not representative of the underlying population, and the data that is eventually reported has several additional shortcomings. Clearly, the situation is far from ideal. To borrow from the computer science literature, there is no silver bullet [136]—there is seemingly no single development in traditional syndromic surveillance that can provide an order-of-magnitude improvement in surveillance timeliness, resolution, or accuracy. However, in very recent history a new type of "surveillance" has emerged that has the potential to complement and alleviate some of the shortcomings of our current surveillance system—*digital surveillance*.

### 4.1.1 The gold standard is not ground truth

As discussed briefly in Chapter 1, weighted percent influenza-like illness (wILI)—reported voluntarily though the ILINet program and posted publicly by CDC—is currently the gold standard of flu activity in the US. Unfortunately, this signal, by the fault of no specific party, has a large number of issues which preclude its use in real-time tracking of localized flu incidence in the US. Some of these are due to the inherent nature of infectious disease. Examples include the fact that different types and subtypes of influenza have varying severity in clinical presentation, access to healthcare is nonuniform, there is an incubation period during which a person is infectious before symptoms appear, and severity of symptoms depends on age and overall health status. Other issues are due to man-made restrictions in the interest of privacy and include the fact that the finest geographic resolution at which wILI is publicly available is at the level of HHS or census regions. Finally, there are issues that are inherent to the data collection process. This is the source of the inherent lag between illness and reporting and is the reason that backfill of provider reports causes retrospective adjustments to previously published wILI.

Revisions due to backfill have the potential to meaningfully change the story of an epidemic, and the magnitude of the wILI update appears to be largest within the first few weeks. This is illustrated for HHS region 9 in Figure 4.1 by overlaying the initial wILI signal (a one week delay) with the final wILI signal (a one *year* delay); an accurate real-time estimate does not exist. Similar—though perhaps not as egregious—revisions are observed across all regions and for the US as a whole. In general, I define the final wILI signal as the value of wILI as reported 52 weeks after the week in question, and I often call this "final" wILI. All other wILI reports I refer to as "preliminary", particularly when referring to the most recently published 1–5 wILI values which are especially volatile.

Some of these issues are unavoidable; for example, there is no plausible way to detect infection before the appearance of symptoms, vulnerable populations (for example, the elderly) will generally present more severe symptoms, and some types of influenza (for example, subtype A/H3N2) are more clinically significant than other types (for example, type B). Other issues, however, might not be insurmountable—at least in theory. The best example of this is the geographic restriction currently imposed on wILI reporting. Each report comes from a healthcare provider that has, at least, an associated zip code—but this information is not shared publicly. It seems difficult to improve the timeliness and stability of wILI, but as I will soon demonstrate, there are new and alternative data sources that can help to do just that.

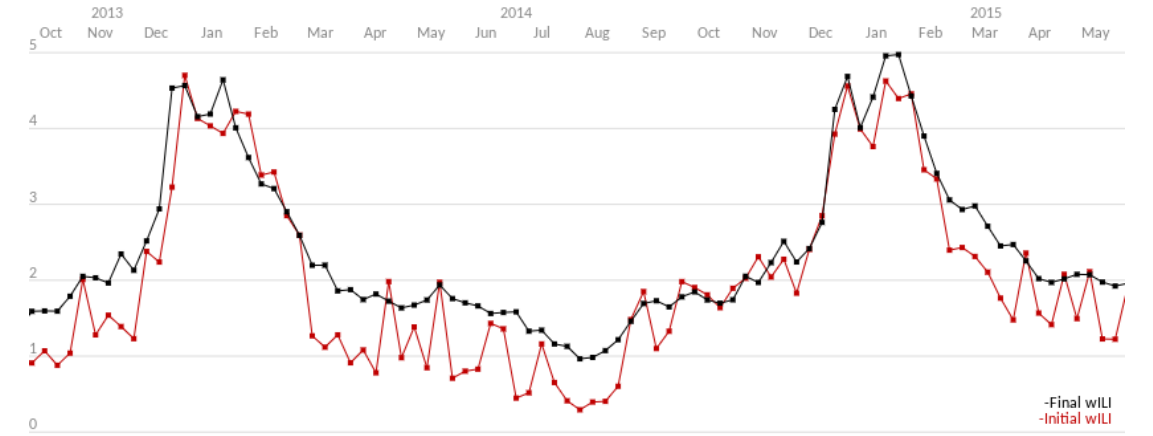


Figure 4.1: **Significant adjustments to wILI are possible through backfill.** In HHS region 9, initial wILI reports (red) are compared with final wILI reports (black).

### 4.1.2 The rise of digital surveillance

The explosive growth of the internet and of the number of individuals with internet access has had a profound impact on society and culture in the past two decades. The most popular websites generate a staggering amount of traffic, with an average hit rate on the order of several hundred thousand hits per *second*. The data, and the statistics based on it, reveal much about current events and about human behavior in general.

Suppose someone develops a fever and sore throat, and, wondering what could cause such a condition, he or she queries an internet search engine (or maybe just asks a smartphone): "Do I have the flu?". This *exact* scenario—word for word—happens *en masse* every flu season; and Google, with some caveats, makes this data available in real-time for various locations within the US (Figure 4.2). There is a clear correlation between what people are searching for and the status of flu epidemics. This discovery led to the creation of a tool built specifically for the purpose of estimating wILI using search queries: Google Flu Trends [69].

Internet users do much more than just submit queries to search engines, and there are two additional types of internet activity that have been shown to be strongly correlated with flu outbreaks. The first of these is reports of self-classified illness posted publicly to social media websites like Twitter. In addition to indicating that a user has been infected, these tweets often contain geolocation metadata which enables us to estimate rates of infection within specific locations. The second activity I broadly refer to as "information-seeking behavior" and includes visits to informational resources like Wikipedia and the CDC website. Both of these sites maintain access logs for the purpose of analytics, and these logs tell a story about what kind of information people are looking for during the flu season. Wikipedia access logs are publicly available and have been used in the past to estimate flu activity in the US [82]; CDC website access logs are not publicly available, but CDC has shared some of these logs with me for research purposes.

Each of the above examples can be used to produce an estimate, or a proxy, of wILI. What is especially appealing about all of these digital surveillance sources is that they are available online and in real-time. Many of them are also available at a finer geographic resolution than

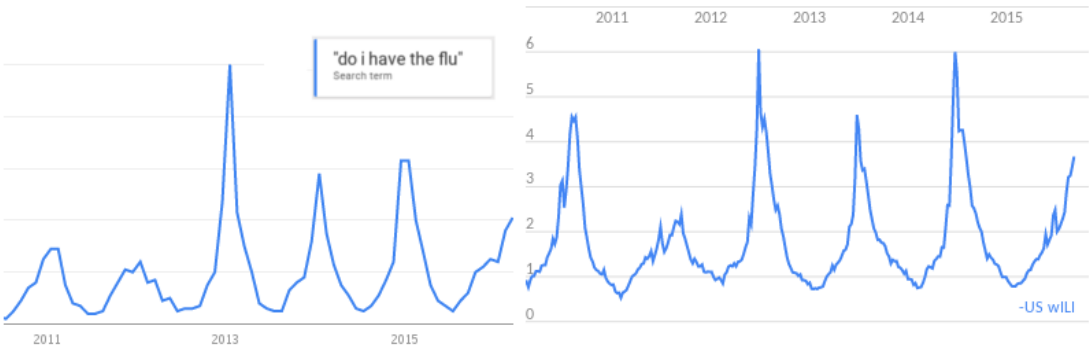


Figure 4.2: **Google Trends data, an example of digital surveillance, resembles wILI in the US.** Left: a screenshot of the Google Trends result for "do i have the flu" between 2011 and 2016. Right: US wILI over the same period.

publicly available ILINet data. The present challenge is to combine all of the signals available at any given point in time to produce an optimal estimate of the final value of wILI at as fine of a geographic resolution as possible. This is in essence a problem of *data assimilation*.

## 4.2 A strategy for optimal assimilation

### 4.2.1 The Kalman filter

The Kalman filter (KF) is an iterative algorithm for estimating the state of a system using **a)** a model of state evolution and **b)** a set of measurements of observable system properties [137]. Perhaps the single most attractive property of the KF is that the mean estimate of the state will be optimal in a least-squares sense—assuming the process (both in reality and as modeled) is linear and all distributions are Gaussian. Additionally, the estimated state is not just a point prediction, but is instead a multivariate normal distribution over state-space. Another nice property of the KF is that it represents a Markov process—the state at time $t$ depends only on the state at time $t-1$. This implies that the KF is recursive and be computed "on-line" without the need to store all past observations. Finally, the KF is especially robust to measurement noise; the observation of *any* new measurement results in a reduction in the uncertainty of the state of the system. The KF operates by repeating the following two steps:

**Predict** Estimate the next state of the system, given only an estimate of the current state.

**Update** Correct the state estimate by incorporating measurements of system properties.

In other words, the KF does the following two things: it predicts what the next state will be, and then it corrects that prediction based on measurements. At all times the system state and measurements are modeled using multivariate normal distributions. Initially ($t = 0$), the KF begins with a user-supplied prior. In the predict step, the prior is one of two distributions that are *added* to form the new state estimate; the other distribution is the predicted change in state that comes from the model of the process. In the update step, the intermediate prior is *multiplied* with

a pre-fused distribution based on measured properties of the system. The resulting distribution becomes the new prior in the next iteration, and the process continues each time a new set of measurements becomes available.

For all its benefits, the original KF is ill-suited for tracking a flu signal like wILI for a couple of reasons. First, the epidemic process is nonlinear—in fact, it is a serious challenge even to understand and to describe this process (see, for example, Chapter 3). Second, there is no canonical model of the process; instead, there are a large number of competing models of flu epidemics, each having unique strengths and weaknesses (see, for example, Chapter 5).

Ignoring these issues momentarily, suppose that we completely give up on modeling flu. As previously mentioned, there exists a large number of digital surveillance streams that can give an indication of the state of the flu epidemic. It would be trivially possible to use the KF with a "no-op" process (for example, predict that all states are equally likely) and still optimally assimilate the available data streams. While a step in the right direction, this is, of course, less than desirable because it leaves out valuable information about how the state evolves.

Taking the example one step further, suppose that instead of using a process model we predict the state update and treat it as if it were a measurement. This strategy allows us to incorporate expert knowledge of the epidemic process from *any number* of flu models while simultaneously producing an optimal estimate of the state based on available measurements. In essence, we have reduced the task of *filtering* to one of *fusion*. The KF inherently performs data fusion as part of the update step, and I exploit this to derive a sensor fusion kernel in the next section. This allows me to fuse any number of predictions and measurements—treating both types of input as *sensors*—into an optimal (up to the assumptions previously discussed) estimate of the state of the epidemic.

### 4.2.2 Derivation of the sensor fusion kernel

As previously discussed, the KF is a two-step process. These steps can be more precisely defined in matrix notation which I will explain shortly. The *predict* step is:

$$\mathbf{x}_{t|t-1} = \mathbf{F}_t \mathbf{x}_{t-1|t-1} + \mathbf{B}_t \mathbf{u}_t, \tag{4.1}$$

$$\mathbf{P}_{t|t-1} = \mathbf{F}_t \mathbf{P}_{t-1|t-1} \mathbf{F}_t^T + \mathbf{Q}_t. \tag{4.2}$$

And the *update* step is:

$$\mathbf{S}_t = \mathbf{H}_t \mathbf{P}_{t|t-1} \mathbf{H}_t^T + \mathbf{R}_t, \tag{4.3}$$

$$\mathbf{K}_t = \mathbf{P}_{t|t-1} \mathbf{H}_t^T \mathbf{S}_t^{-1}, \tag{4.4}$$

$$\mathbf{y}_t = \mathbf{z}_t - \mathbf{H}_t \mathbf{x}_{t|t-1}, \tag{4.5}$$

$$\mathbf{x}_{t|t} = \mathbf{x}_{t|t-1} + \mathbf{K}_t \mathbf{y}_t, \tag{4.6}$$

$$\mathbf{P}_{t|t} = (I - \mathbf{K}_t \mathbf{H}_t) \mathbf{P}_{t|t-1}. \tag{4.7}$$

These equations require some explanation. Throughout, $t$ is a time step—an integer that counts the number of iterations. Any object created at time $t$ is given a subscript of $t$; objects created during the previous iteration are given a subscript of $t-1$. $\mathbf{x}$ and $\mathbf{P}$ are the location and

scale parameters (mean and covariance, respectively) of the multivariate normal distribution that describes the state of the system. From here on, I use the term "state" to refer to $\mathbf{x}$ and "state covariance" to refer to $\mathbf{P}$.

In the *predict* equations, $\mathbf{F}$ is the process matrix, $\mathbf{B}$ is the control matrix, $\mathbf{u}$ is the vector of control inputs, and $\mathbf{Q}$ is the covariance matrix of process (and control) noise. Equation 4.1 predicts the current state of the system, given the last state estimate. It applies the process matrix to the most recent estimate of the state and adds the expected state change due to control inputs. Although the control model ($\mathbf{B}$) and inputs ($\mathbf{u}$) are not used in this analysis, they could be used, for example, to model the expected effects of intervention strategies. Equation 4.2 modifies the state covariance to reflect a decrease in certainty (increase of covariance) in the system state due to noise in the process and control models. This is achieved by projecting the most recent state covariance through the process model and adding process noise.

In the *update* equations, $\mathbf{H}$ is a matrix that maps state space onto measurement space, $\mathbf{R}$ is the covariance matrix of measurement noise, and $\mathbf{z}$ is the vector of measurements. Several temporary variables are created before updating $\mathbf{x}$ and $\mathbf{P}$. The first of these, $\mathbf{S}$, represents uncertainty (covariance) in measurement space (Equation 4.3). It is calculated by projecting state covariance into measurement space and adding measurement covariance. The next, $\mathbf{K}$, is the Kalman gain (Equation 4.4). This represents the relative degree of confidence in the state inferred from the measurement as opposed to the previously predicted state, and it is therefore a function of the predicted state covariance and the covariance of the state after adding measurement noise. Intuitively, it specifies the optimum combination of the predicted state and the measured state, based on the uncertainty of each. When the measurement is relatively more noisy than the prediction, the Kalman gain will favor the prediction; and *vice versa*. The last of the temporary variables is $\mathbf{y}$, the difference between the measurement and the state projected into measurement space (Equation 4.5). Finally, in Equation 4.6 and Equation 4.7 the state and its covariance are updated by an amount determined by the Kalman gain. The new state is a mixture of the predicted state and the measured state, and the new covariance is a scaled-down version of the predicted covariance.

It is important to note the effect that these steps have on the state covariance. The *predict* step (Equation 4.2) adds to the covariance; uncertainty is increased. On the other hand, the *update* step (Equation 4.7) scales-down the covariance; uncertainty is decreased. This makes sense intuitively because no new information is assimilated when making a prediction, so there is no way that certainty in the state could increase (assuming an equilibrium process). Similarly, new information, however noisy, provides more evidence for state estimate.

I now derive the sensor fusion kernel from the canonical KF equations. To begin, I attempt to simplify the notation. First, I assume that all equations are conditional on time $t$. This allows me to drop most of the subscripts—keep in mind though that all of the matrices and vectors are time-dependent. Second, I distinguish between $\mathbf{x}_{t|t-1}$ and $\mathbf{x}_{t|t}$ by introducing a simpler alias for each. I use $\hat{\mathbf{x}}$ in place of $\mathbf{x}_{t|t-1}$ and $\mathbf{x}$ in place of $\mathbf{x}_{t|t}$; I use the same convention for $\mathbf{P}$. Third, I rewrite the *update* equations to eliminate temporary variables. Here are the simplified equations:

$$\hat{\mathbf{x}} = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{u}, \tag{4.8}$$

$$\hat{\mathbf{P}} = \mathbf{F}\mathbf{P}_{t-1}\mathbf{F}^T + \mathbf{Q}, \tag{4.9}$$

$$\mathbf{K} = \hat{\mathbf{P}}\mathbf{H}^T(\mathbf{H}\hat{\mathbf{P}}\mathbf{H}^T + \mathbf{R})^{-1}, \tag{4.10}$$

$$\mathbf{x} = \hat{\mathbf{x}} + \mathbf{K}(\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}), \tag{4.11}$$

$$\mathbf{P} = (\mathbf{I} - \mathbf{K}\mathbf{H})\hat{\mathbf{P}}. \tag{4.12}$$

Notice that the KF combines two things: exactly one prediction and any set of measurements. Clearly, a problem arises if no prediction is available. Is it still possible to use the KF in this case? Similarly, it is problematic if there is more than one prediction. Do all but one have to be discarded? Should they be combined beforehand—and how might this be done? A more interesting case is when the line between prediction and measurement is blurred. Consider, for example, an autoregressive (AR) model. AR is a classic method of time series prediction, but it depends on the last several state values. To use an AR model as a prediction, the KF "state" has to explicitly store not just the current state, but also every preceding state, up to the dimension of the AR model. This in some sense violates the Markov property of the KF—the next state should be conditionally independent of all past states, given the current state. On the other hand, it is trivial to treat the output of an AR model as a measurement; but this is wrong on some level because this "measure" contains no new information about the state of the system—it is a function only of past states. The AR model certainly provides useful information, but is it a prediction or a measurement?

I now derive a special case of the KF which is free from the issues discussed above. In doing this, I show that the KF encodes a sensor fusion operation and that the prediction and all measurements are instances of more general *sensors*.

The root of the problem is the prior; I have no canonical process model for flu, so it is impossible to predict subsequent states. I want to use the KF to assimilate all available data, but it is unclear in the classic formulation above how one might do this when the process is undefined. To proceed, I need to put the KF equations into a workable form. I do this by rewriting the Kalman Gain term using a matrix identity due to [138, 139] and by rewriting the state covariance in so-called "Joseph form" [140]:

$$\mathbf{K} = (\hat{\mathbf{P}}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}, \tag{4.13}$$

$$\mathbf{P} = (\mathbf{I} - \mathbf{K}\mathbf{H})\hat{\mathbf{P}}(\mathbf{I} - \mathbf{K}\mathbf{H})^T + \mathbf{K}\mathbf{R}\mathbf{K}^T. \tag{4.14}$$

Before proceeding further, a discussion on what it means for a process to be undefined is needed. If the process model has no role in updating the state of the system, then the process is not sequential. It is instead a one-step procedure that always starts from the same prior distribution. Further, I make this prior as uninformative as possible by setting it to the uniform pseudo-likelihood over state space—a multivariate Gaussian with infinite variances. Intuitively, without knowing how the process works, all outcomes become equally believable. Mathematically, this suggests that the uncertainty in the estimated state approaches infinity. As the covariance of a

multivariate normal distribution approaches infinity (or, as the precision approaches zero), the distribution effectively becomes an unbounded uniform distribution—a flat improper prior. Following the example of [140], I define the situation of an unknown process model simply by letting $\hat{\mathbf{P}} = \infty$, or equivalently $\hat{\mathbf{P}}^{-1} = 0$. This follows naturally from Equation 4.9 when the process covariance $\mathbf{Q}$ (uncertainty in the state due to noise in the process) approaches infinity. Plugging this into Equation 4.13 gives:

$$\mathbf{K} = (\hat{\mathbf{P}}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}, \tag{4.15}$$
$$\mathbf{K} = (0 + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}, \tag{4.16}$$
$$\mathbf{K} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}. \tag{4.17}$$

Equation 4.14 is more problematic as $(\mathbf{I} - \mathbf{KH})\hat{\mathbf{P}} = 0 \cdot \infty$. To handle this problem, I rewrite Equation 4.14 using the properties of vector covariance:

$$\mathbf{P} = (\mathbf{I} - \mathbf{KH})\hat{\mathbf{P}}(\mathbf{I} - \mathbf{KH})^T + \mathbf{KRK}^T, \tag{4.18}$$
$$\mathbf{P} = (\mathbf{I} - \mathbf{KH})\text{cov}(\mathbf{x} - \hat{\mathbf{x}})(\mathbf{I} - \mathbf{KH})^T + \mathbf{KRK}^T, \tag{4.19}$$
$$\mathbf{P} = \text{cov}((\mathbf{I} - \mathbf{KH}) \cdot (\mathbf{x} - \hat{\mathbf{x}})) + \mathbf{KRK}^T, \tag{4.20}$$
$$\mathbf{P} = \text{cov}((\mathbf{I} - (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}) \cdot (\mathbf{x} - \hat{\mathbf{x}})) + \mathbf{KRK}^T, \tag{4.21}$$
$$\mathbf{P} = \text{cov}((\mathbf{I} - \mathbf{I}) \cdot (\mathbf{x} - \hat{\mathbf{x}})) + \mathbf{KRK}^T, \tag{4.22}$$
$$\mathbf{P} = \text{cov}(0 \cdot (\mathbf{x} - \hat{\mathbf{x}})) + \mathbf{KRK}^T, \tag{4.23}$$
$$\mathbf{P} = \mathbf{KRK}^T, \tag{4.24}$$
$$\mathbf{P} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{R}\mathbf{R}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}, \tag{4.25}$$
$$\mathbf{P} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}, \tag{4.26}$$
$$\mathbf{P} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}. \tag{4.27}$$

Next, I plug Equation 4.17 and Equation 4.27 into Equation 4.11:

$$\mathbf{x} = \hat{\mathbf{x}} + \mathbf{K}(\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}), \tag{4.28}$$
$$\mathbf{x} = \hat{\mathbf{x}} + \mathbf{Kz} - \mathbf{KH}\hat{\mathbf{x}}, \tag{4.29}$$
$$\mathbf{x} = \hat{\mathbf{x}} + \mathbf{Kz} - (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\hat{\mathbf{x}}, \tag{4.30}$$
$$\mathbf{x} = \hat{\mathbf{x}} + \mathbf{Kz} - \hat{\mathbf{x}}, \tag{4.31}$$
$$\mathbf{x} = \mathbf{Kz}, \tag{4.32}$$
$$\mathbf{x} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{z}. \tag{4.33}$$

At last, combining Equation 4.27 and Equation 4.33, we come to what I call the sensor fusion kernel of the KF (in agreement with [140]):

$$\mathbf{P} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}, \tag{4.34}$$
$$\mathbf{x} = \mathbf{PH}^T\mathbf{R}^{-1}\mathbf{z}. \tag{4.35}$$

At this point it may be helpful to zoom out and look at the high-level picture in Equation 4.34 and Equation 4.35. What this says is that, even *without prior knowledge*, it is still possible to produce an optimal estimate of the state of the system using only information from the available measurements. As expected, this is a one-step procedure, and there is no reference to any prior distribution. It may be helpful to draw a parallel to the formula for multiplying univariate normal probability density functions (PDF):

$$\sigma^2 = \left( \sum_i^N \frac{1}{\sigma_i^2} \right)^{-1},$$
$$\mu = \sigma^2 \sum_i^N \frac{\mu_i}{\sigma_i^2}.$$

The KF—and sensor fusion—performs the normal PDF multiplication while additionally accounting for both the correlation structure of the measurements and the mapping from measurement space to state space. Additionally, it can be shown that under specific conditions ($\mathbf{H} = 1, \sum \beta = 1$), sensor fusion is mathematically equivalent to multiple linear regression. The advantage of sensor fusion lies in the descriptive power of $\mathbf{H}$ in mapping between state space and measurement space.

## 4.3 Proxies of flu activity in the US

The biggest challenge in nowcasting is to acquire, in real-time, reliable *measurements* of influenza activity—signals based on digital surveillance. Complementary to these signals are *predictions*—signals that model the epidemic. Whereas measurements (digital surveillance) provide new information about the world, predictions (models) only operate on lagged data. However, unlike measurements, predictions (potentially) contain valuable information in the form of knowledge of the epidemic process. For example, we know that flu epidemics should generally have a single, well-defined peak sometime between December and March. Unlike signals based on digital surveillance, signals based on predictions are "process-aware", and as such will provide a reasonable prediction of the state of the epidemic that is based exclusively on past data. In recent years, many such signals have come and gone. I provide a survey below of the signals that have been used to predict influenza. Of course, the distinction between measurements and predictions is not always so clear, and this is one of the main motivations for using source-agnostic sensor fusion instead of taking a more traditional KF approach.

### 4.3.1 Measurements

#### Google Flu Trends (GFT)

GFT was a very popular indicator of flu activity based on search queries. The model was originally released in 2008 (with retrospective predictions going back to 2003) and was subsequently updated in 2009, 2013, and 2014 [69]. After attracting a great deal of attention from both

academia and popular media, Google discontinued GFT just before the start of the 2015–2016 flu season [72, 141, 142, 143].

This signal is somewhat unique among the digital surveillance sources in that it attempts to *directly* infer current wILI. The dataset contains in-sample, or at least retrospective, predictions from 2003w40 through around 2008w53. Out of sample predictions made in real-time are available from around 2009w01 until 2015w32—the week GFT was discontinued. Predictions are available for many countries, including the United States. Within the US, predictions are available for all states, including the District of Columbia (DC), and around one hundred large cities. Although this dataset is no longer being updated, it is still publicly and freely available from Google [144].

### Google Trends (GT)

GT is a service for exploring the relative popularity of search terms. It is unrelated to GFT, other than that they are both based on search query data. There are several reasons why it is not possible to reconstruct GFT using GT data. First and foremost, there is no official application programming interface (API) for GT; any programmatic access to the service is likely against the terms of service. Second, popularity is measured on an arbitrary scale that varies between locations and time periods. Third, in the interest of privacy, an unspecified threshold is used to hide results for queries with too little volume. Fourth, the terms used to build GFT are not public knowledge, so any attempt to reconstruct GFT would most likely use a somewhat different set of terms.

Even though GT is not a replacement for GFT, it is still a valuable source of data. There have been several attempts to use GT as a signal for nowcasting and forecasting of flu and other infectious diseases [85, 145, 146]. GT is generally available from around 2003 through the present. Temporal and geographic resolution vary greatly depending on search volume. It is possible, with very popular queries, to retrieve weekly values for all US states. GT is publicly and freely available from Google [147], but accessing the data programmatically does not appear to be officially sanctioned.

### Google Health Trends (GHT)

Shortly after Google discontinued GFT, they agreed to support ongoing research in public health and epidemiology by way of a new API—GHT. Although still in development, GHT will one day be a superset of both GFT and GT. Like GFT, the main focus is to provide a query-based signal related to issues of public health, including flu. Like GT, it is possible to query arbitrary search terms across a large number of locations both inside and outside of the US. As a result, GHT opens up the idea of "flu trends" to any number of other diseases or public health issues thanks to the ability to query any set of health-related terms. Like GFT and GT, this signal is available from roughly 2003 through the present.

At the time this thesis work was done, GHT could only provide a weekly signal at the level of US states for a predefined influenza topic. These signals are what I use in subsequent analysis. Current, ongoing extensions of this work are able to benefit from a greater selection of search

terms as more powerful versions of the GHT API have recently been made available by the GHT team.

### HealthTweets (TWTR)

The successes and limitations of GFT inspired the search for new digital surveillance sources, especially those based on social media. Twitter, an immensely popular microblogging service, is an ideal candidate for such a source: *tweets* are incredibly frequent (hundreds of thousands per minute), relatively easy to analyze (no more than 140 characters), often have a clearly-defined topic (via hashtags), are occasionally geotagged (depending on device and preferences), and can be available to the public (depending on preferences). Several attempts have been made to estimate flu activity using Twitter [78, 148, 149]. While these developments were useful in developing the theory of digital surveillance, it was difficult to implement and make use of these ideas in practice.

The Johns Hopkins Social Media and Health Research Group then developed HealthTweets.org, an ongoing, real-time signal of flu activity based on Twitter [79, 84]. This signal, TWTR, is like the GFT of Twitter data in the sense that all of the hard work is already done—what remains is to simply fetch the data. The hard work in this case was a set of tiered classifiers using natural language processing and other machine learning, built with a large set of manually labeled training data. With the recent departure of GFT, TWTR is now one of the best signals available for estimating flu activity in real-time. The data is available from the end of 2011 through the present, at a *daily* resolution, and at the level of US states (and other locations). TWTR is made for the research community, and access is subject to approval and registration [150].

### Wikipedia (WIKI)

Wikipedia, the free, collaborative, and online encyclopedia, is a valuable source of information to people around the world. What enables Wikipedia to be a potential source for digital disease surveillance is that the Wikimedia Foundation makes available the number of visits ("hits"), per hour, to every Wikipedia article. There is a strong correlation between wILI in the US and the number of hits to English articles related to influenza, and several attempts have been made to estimate prevalence of flu (and other infectious diseases) in real-time [81, 82, 83]. Unlike GFT and TWTR, for which most of the manual work has been done and an automatic signal is readily available, there has been no flu signal based on Wikipedia—until now.

To address this void, I created such a signal—WIKI. I intend for WIKI to be for Wikipedia what TWTR is for Twitter and GFT was for Google. WIKI is available weekly from the end of 2007 through the present at the US National level. The WIKI signal is updated weekly on an ongoing basis and is publicly available through the Delphi Epidata API [151].

The methodology of the WIKI signal can be summarized as follows. The number of hourly hits for each Wikipedia article is obtained from `https://dumps.wikimedia.org/other/pagecounts-raw/`. From this, a number of values are extracted. The first is the total number of hits to all English language articles, for subsequent normalization. The rest of the values are the number of hits for each of 54 articles discussed in [81, 83] and enumerated in [151]. These values are stored in a database with metadata including the date and hour during which the hits

were recorded. Whereas previous attempts to create a flu signal from Wikipedia hits aggregated from an hourly to a weekly resolution, I take a slightly different approach. I aggregate to the weekly level *separately* for each hour of the day. The result is a set of weekly values for 54 articles across 24 hours—1296 individual signals. To produce a single estimate of flu prevalence, I first use (in a one-time only, preprocessing step) LASSO regression (using cross validation to select the penalty parameter) to select the articles and hours which are most useful in estimating wILI. I then use the selected signals and unpenalized multiple regression (updated each week) to model the relationship between article hits and wILI as published by CDC.

There are a couple of complications that should be discussed. The first has to do with the size of the training set. I observed that the best predictive power is achieved when some of the training data is excluded. In particular, on any given week the best results were obtained when only training on roughly the most recent year of data. I hypothesize that this is due to somewhat gradual changes in information-seeking behavior, article content, and article notoriety over time. To address this problem, I use a sliding window of 52 weeks to train a separate regression model on each week. Another complication is the fact that values of wILI reported by CDC are subject to subsequent revision, as previously discussed. To prevent preliminary wILI values from reducing model power, I exclude the three most recent wILI values when training the model. A final complication is that the Wikipedia dataset is devoid of any geographic information. It has been common practice to make the assumption that hits to English articles come only from the US, but this is clearly far from the truth. I am able to mitigate this to some extent by selecting articles by hour of day (assuming that the proportion of English hits from the US varies by hour of day), but it remains impossible for now to achieve any finer geographic resolution.

### CDC Page Visits (CDCP)

Given the correlation between Wikipedia page visits and wILI in the US, it seems likely that information-seeking behavior in general can be used to estimate trends in public health. The top links in Google for several flu-related queries lead, of course, to Wikipedia—but also to CDC. CDC collects website analytics for each of their pages, and they have graciously agreed to share some of this data with us for research purposes. To my knowledge, this is a novel—and potentially very valuable—data source, having never before been used to estimate flu prevalence.

The raw data consists of the number of hits, per page *title*, per US state, per day. My goal is to produce a unified signal from this data: a weekly estimate of %ILI within each US state. While the data is available from the start of 2013, as of this writing the total number of distinct page titles is 6,644. As with WIKI, it is necessary to select a subset of the most informative pages. To do this, I used a series of heuristics to select a reasonable set of page titles for use in subsequent modeling. I first limited my search to only the top 50 (arbitrarily) page titles by total hits, reasoning that the signal-to-noise ratio of low-volume pages would likely be lower than that of high-volume pages. Next, I examined the selected page titles and manually excluded those for which I assumed traffic was not primarily driven by having flu symptoms (for example, excluding pages specific to vaccination). With the roughly 15 remaining pages, I excluded those with incomplete time-series (presumably having not been created until sometime after 2013). This exercise led me to discover that page titles generally evolve in minor ways over time, and

so I further limited my search to pages for which I could trace title changes over the entire time period of data availability. At the end of this process, I selected the following 8 pages for further analysis: "What You Should Know for the Influenza Season", "What To Do If You Get Sick", "Flu Symptoms & Severity", "How Flu Spreads", "What You Should Know About Flu Antiviral Drugs", "Weekly US Map", "Basics", and "Flu Activity & Surveillance". By visual inspection, I noticed that hits to these pages, fit to wILI by linear regression, generally appear to over-shoot wILI during weeks of peak activity. To mitigate this effect, I take the log transform of the counts before further processing. Important directions of future work will be to select pages in a more principled manner and to explore other methods of fitting page hit counts to %ILI. For the current analysis, I aggregate page hits from daily to weekly resolution and then use multiple linear regression to fit page counts to %ILI for each US state.

#### Electronic Health Records (EHR)

EHRs have the potential to be the single best real-time indicator of disease incidence [152]; they are created with little to no lag, are not subject to backfill, exist at incredibly fine geographic resolution (zip code at worst), can be a reliable sample of the population, and reflect clinical diagnosis more accurately than syndromic data. Unfortunately, these datasets are often proprietary and are guarded as closely as trade secrets. As an example of the usefulness of EHR data, the Aetnahealth dataset (provided by the Aetna health insurance company) has been used to *retrospectively* estimate real-time flu prevalence in the US [85]. I list this type of signal in the interest of completeness, but I do not currently have access to any EHR data.

#### Flu Near You (FNY)

FNY is an example of participatory surveillance whereby users volunteer once per week to report whether they (or others in their households) personally experienced flu-like symptoms in the preceding week [153, 154]. The reports of ILI collected by FNY have been shown to be strongly correlated with wILI as reported by CDC [155, 156], and this data has been used to estimate the real-time prevalence of flu in the US [85]. The data is not publicly available, but it appears to exist at a weekly level from around 2012 through the present at the level of zip codes (and other locations, for example Canada). As is the case with EHRs, I do not currently have access to this dataset, and I mention it here only for the sake of completeness.

### 4.3.2 Predictions

#### Epicast (EPIC)

Epicast (discussed extensively in Subsection 5.3.2), is a flu forecasting system that I built which is based on collective human judgment [157]. As such, it is a participatory signal, but unlike FNY, it is not based on participants having flu-like symptoms. Instead, it is the collective expectation of wILI, given tentative wILI on past weeks and wILI of past seasons. EPIC is not surveillance *per se*, but more of a measure of anticipation—which, as I show, can be quite accurate. EPIC is a novel data source that is available weekly from 2014w42–2015w20 and 2015w42–2016w20 for the US nationally and for all HHS regions.

**Seasonal Autoregression (SAR3)**

SAR3 is a regression model that, given the current week number and preliminary wILI values of the past three weeks, provides an estimate of the final wILI value of the current week. In other words, there are two sources of information: the current week number and recent wILI values. Using only the wILI values results in a more general model known as "AR(3)"—an autoregression model that uses the three most recent values. Autoregression is an extremely general framework for time series forecasting and as such has seen application in a wide range of settings including epidemic forecasting. The current week number, however, provides additional information to the model. Because flu in the US is a seasonal occurrence, timing information can be very helpful in predicting flu prevalence. SAR3 is available for the same time period and at the same temporal and geographic resolution as wILI—ongoing weekly for the US nationally and for all HHS and census regions. I originally designed and implemented SAR3 as a simple baseline. Here it serves a different purpose; it provides a reasonable estimate of current wILI given only the current week number and past wILI.

Since it may not be immediately obvious how to incorporate timing information into the model, I now give an overview of the SAR3 methodology. Each year consists of 52 (or, once every 5–6 years, 53) weeks, which I refer to as "epiweeks". The absolute week number is not a good candidate for a time signal for a couple of reasons. Foremost among these is the large discontinuity from week 52 to week 1 that usually falls, unfortunately, in the middle of each flu season. Intuitively, weeks 52 and 1 are similar, epidemiologically; but in a linear model, weeks 52 and 1 are more dissimilar than any other pair of weeks. It is possible to use an adjusted week numbering system—say, for example, that weeks 1–20 are instead called weeks 53–72. Similarly, time can be represented by year and fractional week, like $2016\frac{10}{52}$. This scheme, while certainly an improvement, is still less than ideal because now a pre-season week is encoded with a value far away from a post-season week. On the surface this may sound reasonable, but recall that pre- and post-season weeks (e.g. 2015w30 and 2016w29) are characterized by very little flu activity and are therefore epidemiologically similar. What we really want is a way to encode time such that all pairs of adjacent weeks are assigned a small distance. I propose the following solution: treat week number as an *angle*, and use a sinusoidal description of that angle as the measure of timing. More concretely, I produce the following two predictor variables for each epiweek, $w$: $x_1(w) = \sin(2\pi\frac{w}{N}); x_2(w) = \cos(2\pi\frac{w}{N})$, where $N$ is either 52 or 53, depending on the year. This strategy of including pairs of sinusoidal covariates is known more generally as harmonic regression.

There is one more complication that the SAR3 model attempts to account for: the so-called "holiday effect". This effect is manifested as an aberrant increase in reported wILI from the middle of December through early January, presumably due to a change in healthcare-seeking behavior on these weeks (for example, a reduction in the number of reporting providers and in the number of non-urgent, and thus non-ILI, office visits). Neither season-wide timing nor recent wILI predictors can capture this effect well, so I additionally include a set of indicator variables for weeks 50–1 (or weeks 51–1 for years with 53 weeks); each takes a value of 1 on its assigned week and 0 on all other weeks. These week-specific variables allow the model to add a high-resolution offset to wILI during the major holidays in December and January. It is unclear whether the latent process underlying the holiday effect has an additive or multiplicative

(or some other) effect on wILI, but in practice I find that the additive approach of the SAR3 model successfully reduces the error of the estimate.

In total, the SAR3 model consists of 9 predictor variables: three based on recent wILI, two based on season-wide timing, and four based on week-specific timing.

#### The Archetype (ARCH)

The Archetype (discussed in Appendix B), is another system for flu forecasting that I built. Unlike EPIC, it is purely data-driven and requires no human intervention. Therefore, ARCH is not surveillance at all—it is a forecast. ARCH is a novel data source that is available weekly from 2003w40–2016w20 for the US nationally and for all HHS and census regions. However, the ARCH system does not produce forecasts during the off-season, and so no values are reported between epiweeks 21 and 39, inclusive, on all years.

### 4.3.3 Summary

Given all of the above proxies of flu activity (both surveillance-based and prediction-based), the goal is to forecast the final wILI value eventually reported by CDC.

| Signal | Type | Time Period | Resolution | Access |
|---|---|---|---|---|
| GFT | Search query | 2009–2015 | US States | Public |
| GT | Search query | *not used* | *not used* | Restricted |
| GHT | Search query | 2004–Now | US States | Restricted |
| TWTR | Social Media | 2012–Now | US States | Restricted |
| WIKI | Info. Seeking | 2007–Now | National | Public |
| CDC | Info. Seeking | 2013–Now | US States | Restricted |
| EHR | Insurance Claims | *not used* | *not used* | Commercial |
| FNY | Participatory | *not used* | *not used* | Restricted |
| EPIC | Participatory | 2014–2016 | Regional | Public |
| SAR3 | Prediction | 1997–Now | Regional | Public |
| ARCH | Prediction | 2003–Now | Regional | Public |

Table 4.1: **Summary of digital surveillance and forecasting signals.** Time Period and Resolution columns describe *only* the datasets that I was able to use for nowcasting; full datasets may have more extensive coverage and finer granularity.

### 4.3.4 Fitting digital surveillance to (w)ILI

Most of the previously described flu proxies are not pre-fitted to (w)ILI, and those that are often exhibit systematic bias. I adopt the following strategy for producing an estimate of (w)ILI separately for each data source and location. I use weighted (potentially multiple) linear regression to fit the signal to either wILI (for national and regions) or %ILI (for states). Each fitted version, which I call a "sensor", is then used as input to sensor fusion in which it is assumed that all inputs have zero bias and IID Gaussian noise over time. The regression weights were designed with the following goals: (a) samples at or near to the same week of year should be given more weight than samples on more distant weeks, (b) sample weight should fall exponentially as time passes, and (c) very recent samples should be penalized as wILI on these weeks is subject to revision. I encode these objectives in the following function of the number of weeks, $dw$, spanning a previously observed sample and the current week (plotted in Figure 4.3):

$$
\begin{aligned}
\text{year} &= 52.2, \\
a &= \frac{1}{20} + \frac{19}{20} \exp\left(-(\min(dw\%\text{year}, \text{year} - dw\%\text{year})/2)^2\right), \\
b &= 2^{-(dw/\text{year})}, \\
c &= 1 - 2^{-(dw/1)}, \\
\text{weight}(dw) &= a \cdot b \cdot c.
\end{aligned}
$$

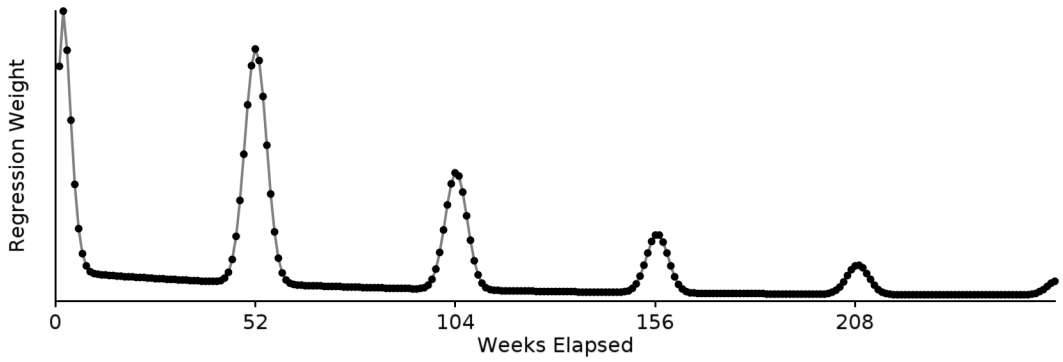


Figure 4.3: **Regression weights for fitting signals to (w)ILI.** Each digital surveillance signal is fit, separately for each location, to either wILI or %ILI (depending on location) using multiple linear regression with weights plotted above. Weights are shown here over a five year period, however the weight vector is generated at runtime with length equal to the number of past observations. Weights are intended to capture seasonal effects, to account for backfill, and to be robust to temporally evolving relationships. Each fitted signal becomes a "sensor"—an approximately unbiased and appropriately scaled estimate of (w)ILI.

Figure 4.4 compares sensor readings of all eight flu proxies to US national wILI from 2011–2016. This highlights some of the main challenges in this data assimilation problem, in particular that sensors are noisy and intermittently available. More difficult to illustrate is that sensors also cover different geographic regions and resolutions, and are likely correlated. Sensor fusion, through careful construction of matrices $\mathbf{H}$ and $\mathbf{R}$, is capable of handling all of these issues.

There is one remaining complication that has not been addressed: while population-weighted %ILI (wILI) is publicly hosted by CDC for the US as a whole and for various regions within the US, %ILI is not publicly available at the level of US states. To fit a given signal to %ILI requires that at least some values of %ILI are known. CDC, through agreement with participating states, has agreed to share %ILI with us from 2002–2015 for 26 states. While this is a fantastic dataset, it is insufficient for fitting signals in the missing states. Fortunately, most states publicly post tables or charts of %ILI online. On 2015w35, I manually visited each of these 51 (including DC) websites and scraped %ILI when it was available. In total, I gathered at least one flu season of

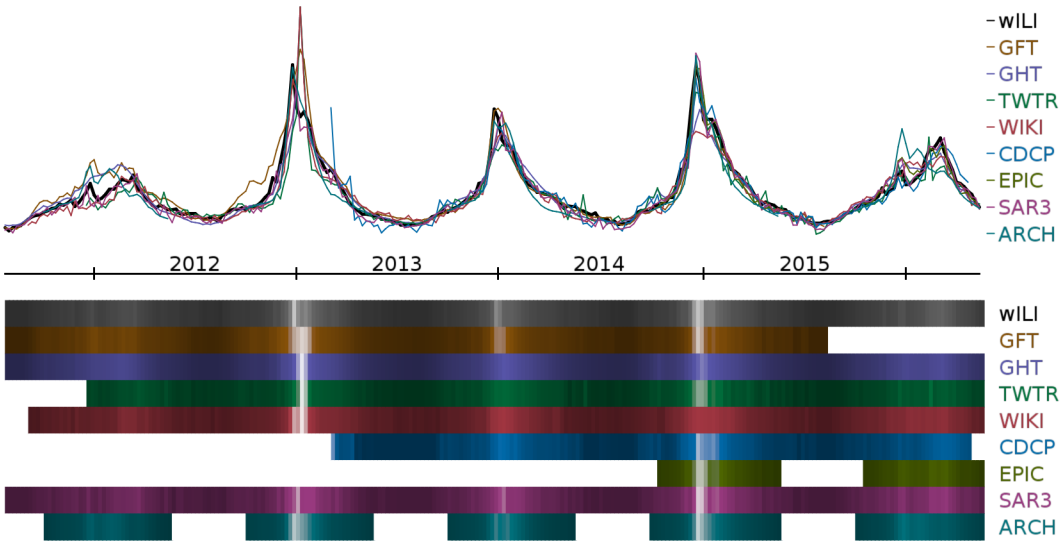


Figure 4.4: **Comparison of "ground truth" (wILI) and sensor readings for the US from 2011–2016.** Both charts depict the same data but highlight different aspects. Top: a line chart of wILI as a function of time, to highlight sensor noise. Bottom: a timeline shaded by signal intensity, to highlight sensor availability.

data for 43 states. Unfortunately, the conversion process is lossy and error-prone, and I found that scraped values do not always exactly agree with official values (most, however, closely agree). Additionally, there are 5 states for which I have neither an official %ILI report nor a scraped %ILI estimate.

I devised a strategy to resolve conflicts between official values and scraped values and to infer %ILI in missing states. The intuition is as follows. First, I strongly trust values shared by CDC and only weakly trust values that I manually scraped. Second, HHS and census regions—for which wILI is known—is a linear combination of %ILI in constituent states, and it is possible to infer missing values for one state when the remaining states in a given region are known. Together, these two ideas led me to an optimization problem: separately for each week, estimate a value of %ILI in each state such that (a) official state %ILI, if available, is very close to estimated %ILI, (b) scraped state %ILI, if available, is at least somewhat close to estimated %ILI, and (c) the population-weighted sum of estimated %ILI in a region is very close to official wILI, for all regions. More precisely, I somewhat arbitrarily penalize squared deviation from official %ILI with weight 1, from scraped %ILI with weight 5, and from regional wILI with weight 1. I use the Nelder-Mead gradient-free optimization method [158] to find an assignment for all state %ILI values given an objective function implementing these penalties. I now have wILI for the US as a whole, wILI for the HHS and census regions, and reasonable estimates of %ILI for all states.

## 4.4 Nowcasting influenza within the US

The primary aim of this chapter is to nowcast (w)ILI in the US. To do this, I use the sensor fusion methodology, digital surveillance signals, and predictive signals described above. In what follows, I show how the design matrices are constructed for the problem at hand, assess accuracy of nowcasts relative to known and estimated wILI and %ILI, and perform sensitivity analysis by selectively withholding portions of the available data.

### 4.4.1 Digital and predictive surveillance in the sensor fusion framework

The sensor fusion kernel in Equation 4.34 and Equation 4.35 makes use of three things: $\mathbf{H}$ to map from state space to measurement space, $\mathbf{R}$ to describe the covariance of sensor noise, and $\mathbf{z}$ which are the sensor readings at any given point in time. Before construction of $\mathbf{H}$, it is necessary to precisely define what is meant in this context by state and measurement spaces. US states, US regions, and the US as a whole form a hierarchy wherein each tier (states, regions, national) is a linear combination of the locations in lower tiers. I define state space to be the finest geographic resolution from which (w)ILI in *all* locations can be calculated: state space is therefore the 51 US states and DC ($\mathbb{R}^{51}$). Note that this definition assumes that at least one sensor is reporting at the level of US states. If this is not the case (for example, perhaps only regional sensors were available at runtime) then state space will be US regions, or in the worse case, the US as a whole. In any case, state space is the lowest tier in the location hierarchy that is available at runtime.

Measurement space is defined by what inputs are available at runtime, the elements of $\mathbf{z}$. As this depends on which data streams happen to be available, it is not possible to give a specific assignment to measurement space in general. Measurement space will be of the form: $(s, l) \forall l \in \text{locations}_s \forall s \in \text{sources}$. In words, measurement space is the set of all source-location pairs, where locations vary by source. For example, on 2015w01 all eight sources were available, and there were a combined total of 308 input sensors ($\mathbb{R}^{308}$).

Up until this point it has been implicitly assumed that the end goal was to estimate state space. In reality, I am interested not only in estimating %ILI in all US states, but also in estimating regional and national wILI. Due to the hierarchal nature of US locations, regional and national wILI is simply a linear combination of %ILI in US states. I now modify the sensor fusion kernel to produce estimates for all locations by introducing a matrix, $\mathbf{W}$, mapping state space to output space (nationally (1), regionally (19), and all states with DC (51); 71 locations in total; $\mathbb{R}^{71}$):

$$\mathbf{S} = \mathbf{W} P \mathbf{W}^T, \tag{4.36}$$

$$\mathbf{S} = \mathbf{W} (\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1} \mathbf{W}^T, \tag{4.37}$$

$$\mathbf{y} = \mathbf{W} \mathbf{x}, \tag{4.38}$$

$$\mathbf{y} = \mathbf{W} \mathbf{P} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{z}, \tag{4.39}$$

$$\mathbf{y} = \mathbf{W} (\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{z}. \tag{4.40}$$

Under this new formulation, $\mathbf{y}$ and $\mathbf{S}$ are together the mean and covariance of a multivariate normal distribution describing estimated (w)ILI in all locations. Therefore, $\mathbf{y}$ provides a point

prediction of (w)ILI for each location, and the diagonal of $\mathbf{S}$ provides the associated variance of each prediction. A graphical overview of the complete sensor fusion process is shown in Figure 4.5.

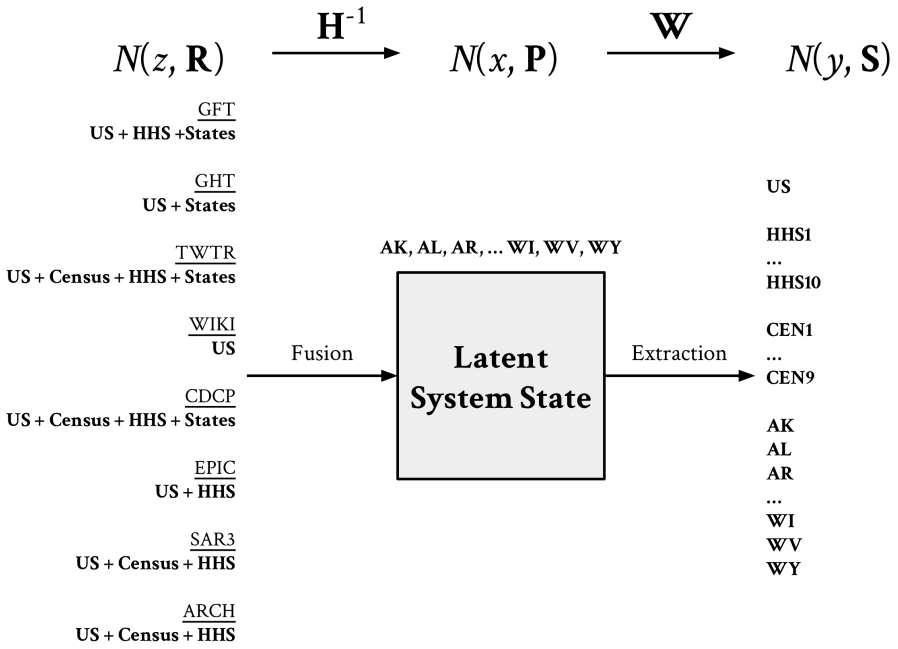


Figure 4.5: **Overview of the sensor fusion and extraction process.** Sensors (left; measurement space) are "fused" to produce an estimate of the latent system state (middle; state space). The desired outputs (right; output space) are then "extracted" from the estimated state. A multivariate normal distribution represents the estimate within each space. Matrices $\mathbf{H}$ and $\mathbf{W}$ map from state space to measurement and output spaces, respectively.

Now it is possible to define $\mathbf{H}$ and $\mathbf{W}$. Both matrices map from state space to another space, with columns corresponding to states ("states" in the sense of state space, which coincidentally in this case is US states). The rows of $\mathbf{H}$ correspond to measurement space (individual sensors), and the rows of $\mathbf{W}$ correspond to output space (all US locations). By nature of the location hierarchy, all rows of $\mathbf{H}$ and $\mathbf{W}$ sum to 1. Additionally, $\mathbf{H}$ has full column rank; in other words, $\mathbf{H}^T\mathbf{H}$ is invertible. (Otherwise some states would be indeterminate.) Finally, the element values of both $\mathbf{H}$ and $\mathbf{W}$ are the fraction of each column's ($j$) population out of each row's ($i$) population, as in:

$$H_{ij} = \frac{\begin{cases}\text{population}_j, & \text{if } j \in i \\ 0, & \text{otherwise}\end{cases}}{\sum_{k \in i} \text{population}_k}.$$

The final task is to produce $\mathbf{R}$, the covariance of sensor noise. It is trivial to estimate this when no data are missing and there are more observations than sensors; in practice, however, sensors are intermittently available and the number of sensors is typically much greater than the number of weekly observations. We are faced with two separate problems: missing data and high-dimensional estimation. A large number of strategies to address these issues are described in the literature, some of which include [159, 160, 161, 162, 163, 164]. Here, for the sake of simplicity and speed, I employ the methods of [165] for handling missing values and [166] for regularization.

By these two methods, I estimate $\mathbf{R}$ as follows. First, I compute the covariance between each pair of sensors over the weeks on which they are simultaneously available. It is important to note that the resulting pairwise covariance matrix, $\hat{\mathbf{R}}$, is not guaranteed to be positive semidefinite (PSD). Second, I find the smallest value of $\alpha \in [0, 1]$ such that $\mathbf{R} = (1 - \alpha)\hat{\mathbf{R}} + \alpha \text{diag}(\hat{\mathbf{R}})$ is PSD. Finally, I shrink the estimate further towards the diagonal matrix of sensor variances as this seems to help in practice, presumably by further avoiding overfitting. To do this, I somewhat arbitrarily set $\beta = \frac{3}{4}\alpha + \frac{1}{4}1$. My final estimate is a blend, with weight $\beta$, between $\hat{\mathbf{R}}$ and $\text{diag}(\hat{\mathbf{R}})$. This matrix is guaranteed to be PSD and invertible, and it has the added benefit of using all available data without imputing missing values.

At last, all components have been defined, and what remains is to run the sensor fusion kernel to estimate (w)ILI in the US. I do this now, ongoing and in real-time, every week. These nowcasts are publicly available via the Delphi Epidata API [151]. Planned future work includes making a public-facing web interface for visualizing these estimates. I have also produced retrospective nowcasts for all US locations from 2011–2016, and below I assess the accuracy of the sensor fusion methodology using these estimates.

### 4.4.2 Results and comparative analysis

Figure 4.6 compares "ground truth" and nowcasts for all locations. I provide this figure not to make a statement about accuracy, but instead to illustrate the final output of the sensor fusion framework for flu nowcasting. Intermittent, noisy, correlated, and geographically distinct signals from a variety of online and offline sources are fused into a unified estimate of flu activity within each location. If the noise of each sensor relative to ground truth was truly zero-mean IID Gaussian, then the nowcast would be optimal in a least-squares sense. This is almost certainly not the case in reality. However, by careful design of the fitting procedures previously described, it is hopefully not too far from the truth.

I use the following metrics to assess the accuracy of nowcasts: Pearson correlation coefficient (PCC; best=1), mean absolute error (MAE; best=0), root mean squared error (RMSE; best=0), and hit rate (HR; best=100). PCC describes correlation, MAE and RMSE describe error, and HR describes the fraction of predictions that change in the same direction as ground truth. Given ground truth, $x$, and nowcasts, $y$, these are defined as:

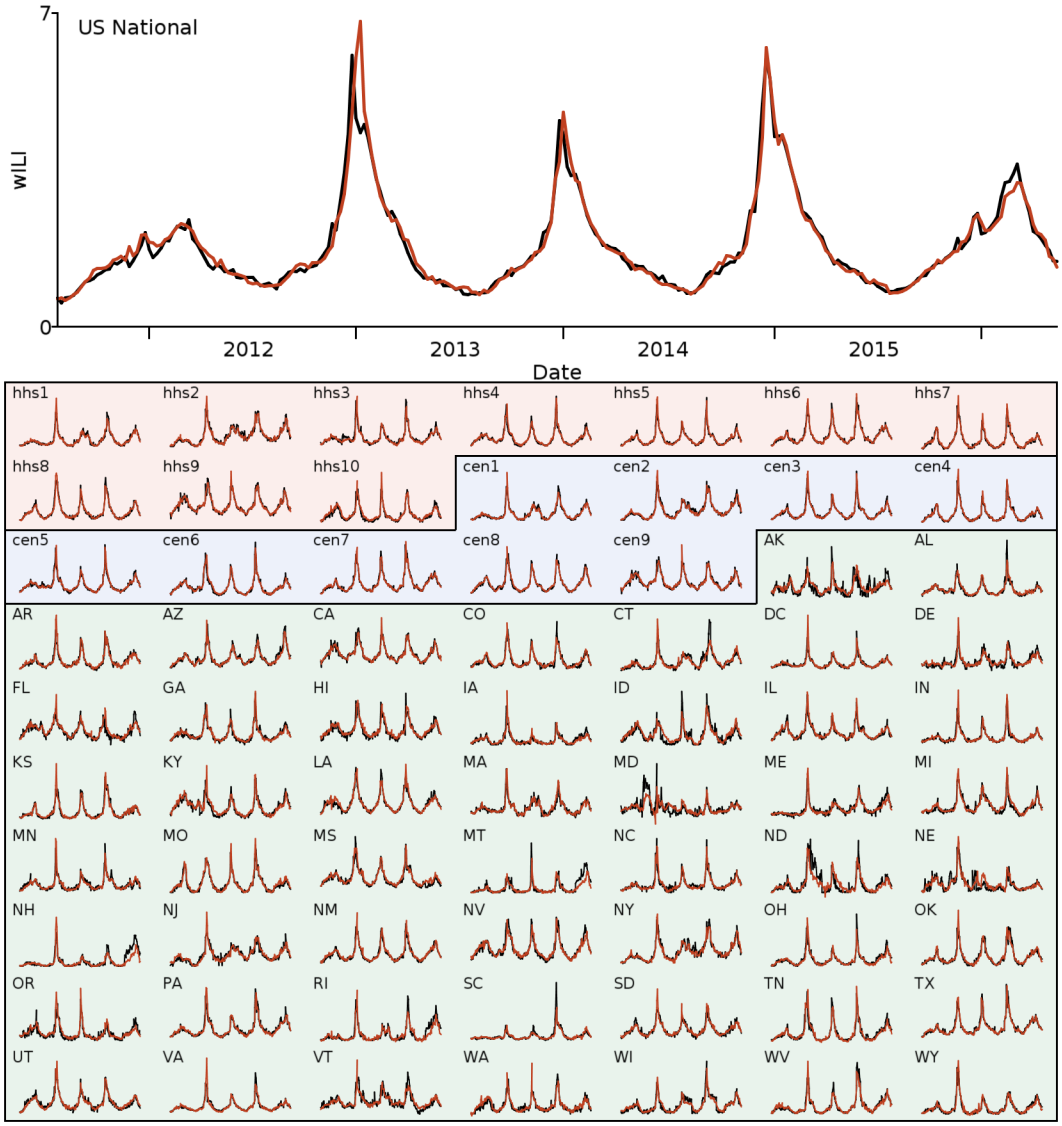


Figure 4.6: **Retrospective out-of-sample nowcasts for all US locations from 2011w30 to 2016w20.** All plots show "ground truth" (either wILI or %ILI, depending on location) in black and corresponding nowcasts in orange. Top: US national. Bottom: HHS regions (shaded light red), Census regions (shaded light blue), and US states (shaded light green).

$$\text{PCC} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|,$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}, 54$$

$$\text{HR} = 100 \times \frac{\sum_{i=2}^{n}\left(\text{sign}(x_i - x_{i-1}) == \text{sign}(y_i - y_{i-1})\right)}{n-1}.$$

I show these metrics, averaged over locations in each tier (national, regional, states), for the full nowcasting period of 2011w30–2016w20 in Table 4.2. Accuracy decreases as geographic resolution increases. This is to be expected for a number of reasons. First, four of the sources (WIKI, EPIC, SAR3, ARCH) are unavailable at the level of US states; all, however, are available nationally. Second, locations in higher tiers are able to incorporate detailed information from locations in lower tiers. For example, wILI in the 10 HHS regions can be used to compute national wILI, but the reverse is not true. Third, because of sampling and other artifacts, (w)ILI is inherently more noisy at finer geographic resolutions and is therefore more unpredictable. Finally, there is a significant caveat to recall at the state level: official %ILI is unknown for 25 states, and for those states I use estimated %ILI as ground truth.

| Tier | # Locs | PCC | MAE (%ILI) | RMSE (%ILI) | HR (%) |
|---|---|---|---|---|---|
| National | 1 | 0.970 | 0.124 | 0.253 | 75.3 |
| Census | 9 | 0.953 | 0.201 | 0.342 | 61.0 |
| HHS | 10 | 0.950 | 0.205 | 0.337 | 60.0 |
| States | 51 | 0.887 | 0.340 | 0.543 | 54.3 |

Table 4.2: **Nowcasting accuracy by location.** Metrics were computed over 252 weeks spanning 2011w30–2016w20.

To contextualize these results, I compare these nowcasts with various other signals and systems at the national level. For fair comparison, I truncate the nowcast in time so that metrics are compared over the same time periods. In what follows, the sensor fusion nowcasting system described above is labeled "Delphi-SF".

First, I compare the nowcast with each of the eight national sensors (Figure 4.7). In other words, I compare the system inputs with the system output. In all of PCC, MAE, and RMSE, Delphi-SF has better performance (higher correlation, lower error) than the eight national sensors. In HR, Delphi-SF is only surpassed by GHT and ARCH. These results serve as a nice sanity check. The output of the system should be, on average, more accurate than the inputs to the system; I find that this is the case.

Next, I compare the nowcast with CDC's first-posted value of wILI ("Prelim.") (Table 4.3). These values are posted on (or after) Friday the week following the observed week. These preliminary values are subject to change over time as more provider reports are collected; this is the backfill effect. Accuracy of the nowcast approaches, but does not exceed, the accuracy of preliminary wILI. However, to put this in perspective, the nowcast is available almost immediately after the observed week has ended—as early as Sunday.

| System | PCC | MAE (%ILI) | RMSE (%ILI) | HR (%) |
|---|---|---|---|---|
| Delphi-SF | 0.970 | 0.124 | 0.253 | 75.3 |
| Prelim. | 0.994 | 0.125 | 0.152 | 76.5 |

Table 4.3: **Comparison of nowcast and preliminary wILI.** Metrics were computed over 252 weeks spanning 2011w30–2016w20.

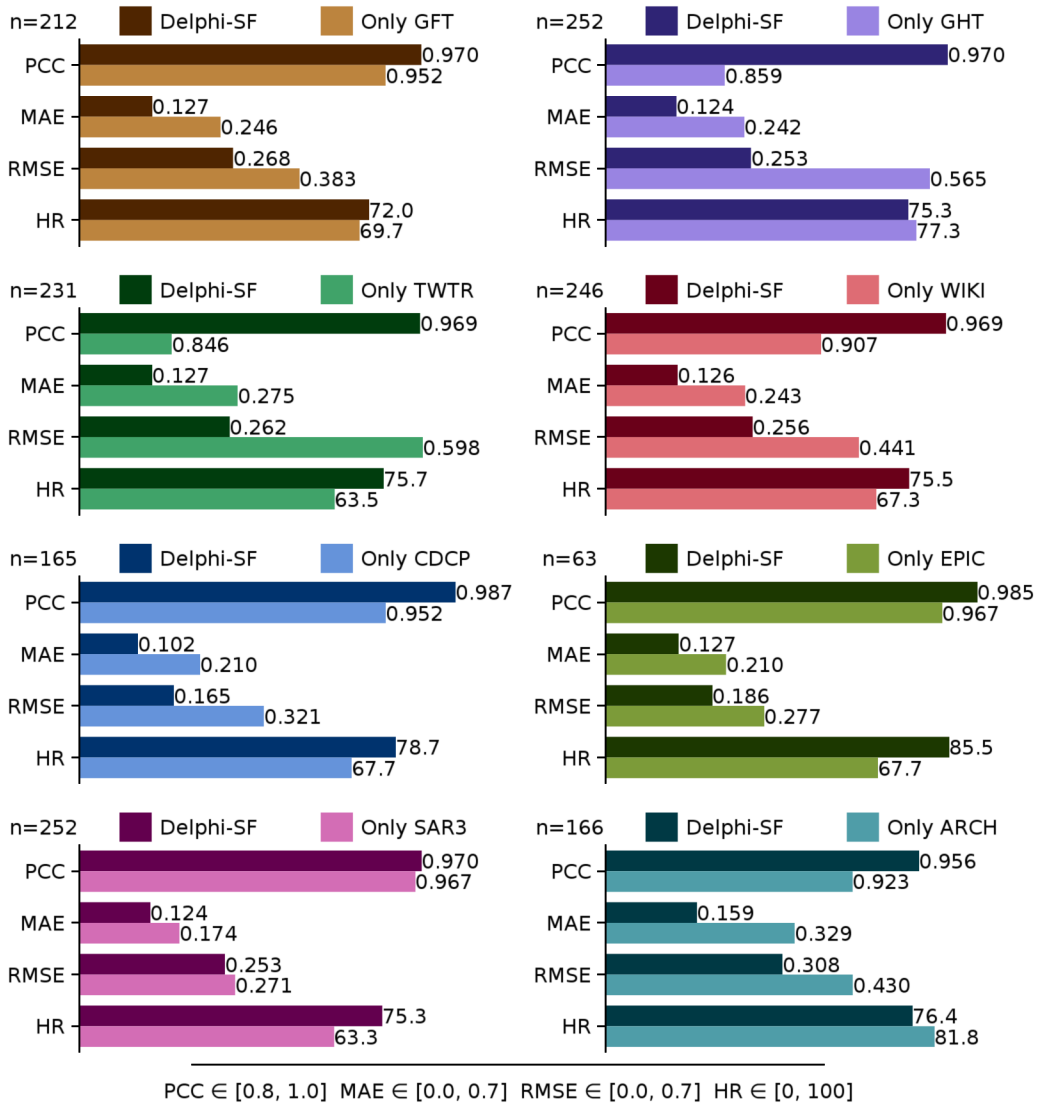


Figure 4.7: **Comparison of sensor fusion system inputs (national sensors) and output (national nowcast).** For each source, metrics for both that source and Delphi-SF were computed over the period of time during which the source was available, as indicated in the top-left corner of each chart. The displayed bounds for each metric are shown at the bottom of the figure. (In PCC and HR, larger is better; in MAE and RMSE, smaller is better.)

Finally, I compare with the "Support Vector Machine regression with Radial Basis Function kernel" (SVM-RBF) nowcasting system of [85] (Table 4.4). They use five digital surveillance signals, all of which I have previously discussed: FNY, EHR, GT, GFT, and TWTR. They produce nowcasts only at the national level, for the period spanning August 2013 to February 2015. It is difficult to compare these systems directly given the significant differences in surveillance

inputs (EHR and FNY) and the small number of locations (one) and seasons (two) for which nowcasts are available. In any case, the accuracy of the two systems appears to be quite similar. The most obvious advantage of Delphi-SF is that estimates are produced at much finer geographic granularities. A more subtle advantage is that Delphi-SF handles missing data, an issue which has not yet been addressed in the SVM-RBF framework.

| System | PCC | MAE (%ILI) | RMSE (%ILI) | HR (%) |
|---|---|---|---|---|
| Delphi-SF | 0.987 | 0.118 | 0.198 | 72.4 |
| SVM-RBF | 0.989 | — | 0.176 | 69.4 |

Table 4.4: **Comparison of nowcasting frameworks.** Metrics were computed over 77 weeks spanning 2013w35–2015w06. SVM-RBF metrics are from [85] (MAE was not reported).

### 4.4.3 Sensitivity analysis

In what follows, I explore the effects of two types of perturbations on nowcasting accuracy. The first is a set of "ablation" experiments in which individual sources are removed. The second is a set of "abscission" experiments in which sensors at individual geographic tiers are removed.

I perform ablation experiments separately for each source as follows. I record all weeks for which the source of interest is available, and only for those weeks I run the sensor fusion framework under two different settings. The first is the standard all-available-sources sensor funsion as before ("Delphi-SF"). The second uses all available sources *except for* the source of interest. In this way it is possible to see the effect of removing each source from the mixure of sensor fusion inputs. The results of these experiments measured nationally are shown in Figure 4.8.

Ideally, withholding any source should not *improve* accuracy; accuracy should either decrease, or remain approximately unchanged. In most cases, I find that this is the case, however there are some exceptions. Removing WIKI, EPIC, SAR3, and ARCH results in uniformly lower accuracy (lower correlation, higher error). Removing GFT, GHT, and CDCP yields mixed results. Removing TWTR results in uniformly higher accuracy; this indicates that the sensor fusion framework is relying too heavily on this source. Improved estimation of the noise covariance matrix $\mathbf{R}$ may help to alleviate this problem.

I perform abscission experiments over geographic tiers as follows. In normal operation, the sensor fusion method takes as input a large set of sensors from the eight sources. These sensors can be classified by one of three geographic tiers: national, regional, and states. In these experiments I withhold the sensors of various tiers to study the effect of input (sensor) resolution on output (nowcast) accuracy, at the national level. I exhaustively test all $2^3 - 1$ combinations of tiers, excluding the degenerate case of no input. For fair comparison, I calculate and assess nowcasts during 2014w42–2015w20 as this is the only period of time in which all eight sources were simultaneously available. The results of these experiments are shown in Table 4.5.

By all measures, the most accurate nowcast was produced when all tiers were enabled ("N+R+S"). This is a good sanity check because, as stated before, witholding data should not *improve* accuracy. By similar reasoning, it is good to see that, in general, accuracy improves when including

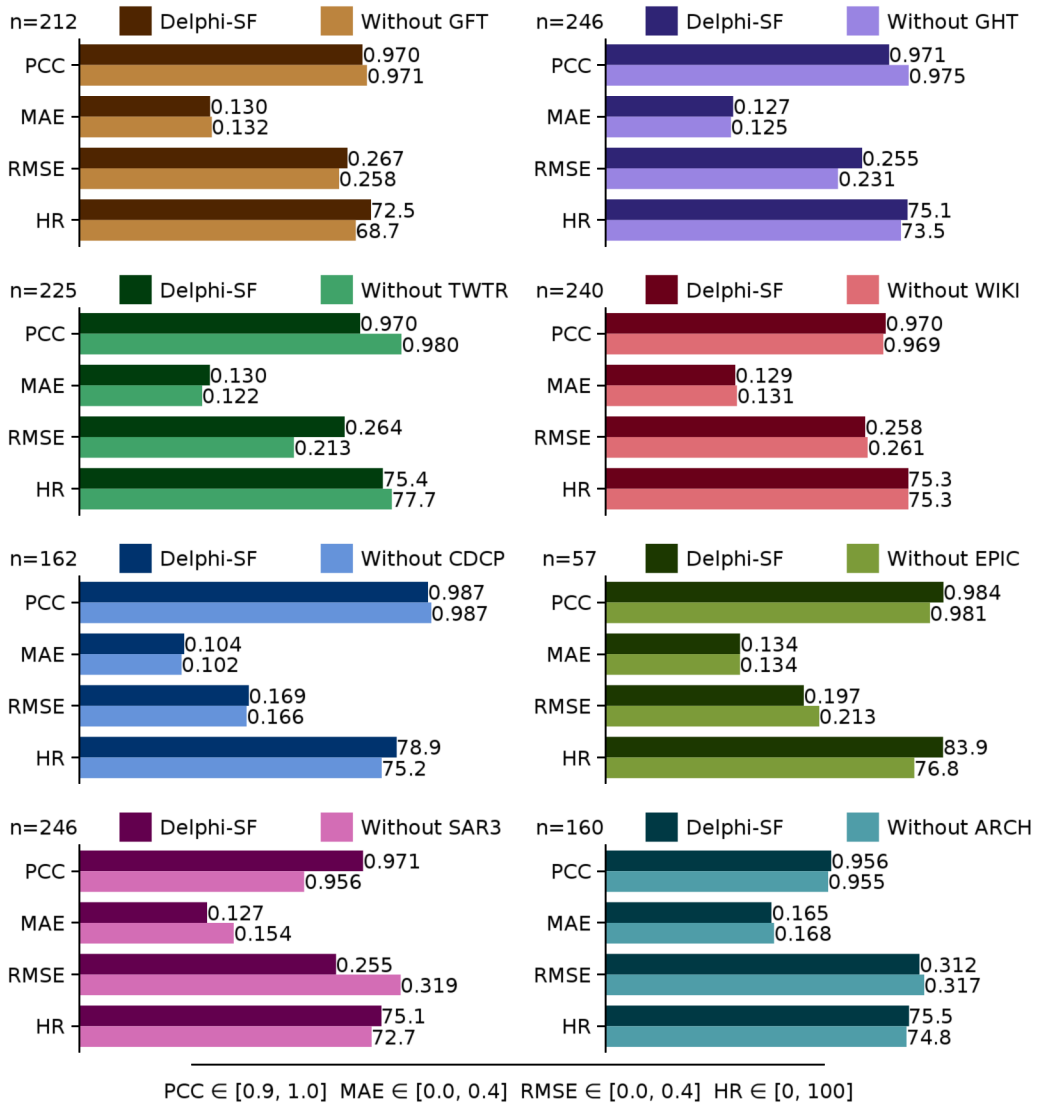


Figure 4.8: **Sensor fusion ablation experiments.** For each source, metrics for Delphi-SF, both with and without that source, were computed over the period of time during which the source was available, as indicated in the top-left corner of each chart. The displayed bounds for each metric are shown at the bottom of the figure. (In PCC and HR, larger is better; in MAE and RMSE, smaller is better.)

sensors in higher resolution tiers. (See: "N" → "N+S", "R" → "R+S", "N" → "N+R", and , "N+R" → "N+R+S".) In comparing individual tiers, the regional tier ("R") is, perhaps surprisingly, more accurate than the national ("N") tier, which is in turn more accurate than the state ("S") tier. It is difficult to interpret the significance of this observation given that the set of sources within each tier differs (recall Table 4.1).

| Input | State | Output | PCC | MAE | RMSE | HR |
|---|---|---|---|---|---|---|
| N | N | N | 0.987 | 0.162 | 0.243 | 80.0 |
| N+R | R | N+R | 0.988 | 0.142 | 0.207 | 76.7 |
| N+R+S | S | N+R+S | 0.988 | 0.137 | 0.203 | 80.0 |
| N+S | S | N+R+S | 0.986 | 0.138 | 0.215 | 80.0 |
| S | S | N+R+S | 0.979 | 0.162 | 0.267 | 73.3 |
| R+S | S | N+R+S | 0.988 | 0.139 | 0.205 | 76.7 |
| R | R | N+R | 0.988 | 0.145 | 0.207 | 73.3 |

Table 4.5: **Sensor fusion abscission experiments with all sources.** Metrics were computed over 31 weeks spanning 2014w42–2015w20. The "Input" and "Output" columns represent the tiers for which measurements and nowcasts are available, respectively. The "State" column represents the geographic resolution of latent state space. "N": US National (1). "R": HHS and census regions (19). "S": US States (51).

To address the issue of having a different set of sources within each geographic tier, I ran additional abscission experiments using only the data sources that are available for all tiers: TWTR and CDCP. These sources are jointly available from roughly 2013w30 through the present. The results of these experiments are shown in Table 4.6. Here I find that accuracy by PCC, MAE, and RMSE increases from lower to higher geographic resolutions. HR is lowest for the national tier and highest for the regional tier. Collectively, these results agree with my intuition that increasing the resolution of the inputs, which has the effect of increasing the resolution at which the latent state space is modeled, should generally increase the accuracy of the output.

| Input | State | Output | PCC | MAE | RMSE | HR |
|---|---|---|---|---|---|---|
| N | N | N | 0.974 | 0.167 | 0.247 | 71.6 |
| R | R | N+R | 0.976 | 0.145 | 0.246 | 73.8 |
| S | S | N+R+S | 0.977 | 0.135 | 0.227 | 72.3 |

Table 4.6: **Sensor fusion abscission experiments with selected sources.** Metrics were computed over 142 weeks spanning 2013w30–2016w15. The "Input" and "Output" columns represent the tiers for which measurements and nowcasts are available, respectively. The "State" column represents the geographic resolution of latent state space. "N": US National (1). "R": HHS and census regions (19). "S": US States (51).

Finally, it may be helpful to contrast these results with those in Table 4.2. When the inputs are fixed and accuracy is measured over tiers of varying resolution, *accuracy increases as resolution decreases*. When the resolution of the input is varied and accuracy is measured over a fixed tier, *accuracy increases as resolution increases*. The difference between these two observations, though subtle, seems to agree with intuition.

## 4.5 Final considerations

Early in this chapter I derived the sensor fusion kernel from the Kalman filter. Although I did not know it at the time, this result was known at least as early as 1997 [140]. What I discovered, it seems, was a different path from start to finish. Along the way, however, I was able to prove the equivalence of special cases of sensor fusion and multiple linear regression, which may be a novel—though somewhat orthogonal to this thesis—result. What is both novel and central to my thesis is the application of the sensor fusion methodology to flu surveillance data for high resolution nowcasting.

While most previous nowcasting attempts have focused on a single digital surveillance stream [66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83], I instead aim to use all available evidence, combining both digital surveillance and short-term forecasts. To accomplish this goal with a modest amount of historical data, I was forced to devise a general fitting strategy for all signals to (w)ILI. There is significant future work to be done in fitting each of these signals to (w)ILI beyond my simple (but reasonable) attempt. In particular, the GHT and CDCP datasets are entirely novel, and there is undoubtedly a large and untapped potential in each of these sources. A closely related task is to optimize the selection of articles for WIKI and pages for CDCP. I made a reasonable attempt at both, but it is important to note that the selections I made (using LASSO and manual inspection for model selection) are likely far from optimal.

Similarly, I present a bona fide attempt to reconstruct historical %ILI at the US state level using a mixture of evidence from a number of official and unofficial sources. However, the problem of unknown %ILI continues going forward in time, and this is an important direction for future work. The ideal solution to this problem is, of course, increased data sharing; however, this may not be possible. In that case, it will be important to either estimate state %ILI or to devise a method for fitting signals to %ILI when %ILI is only available in the past.

Another direction of future work is to improve the estimation of the noise covariance matrix $\mathbf{R}$. The performance of sensor fusion depends critically on having an accurate estimate of this covariance. This is unfortunately a nontrivial task in this setting due to both the intermittency of surveillance inputs and the very high dimensionality of the input compared to the number of weekly observations. One possible approach is to select the $\beta$ blending parameter through an automated process like cross validation. Alternative strategies for estimating the sensor noise covariance (or precision) matrix may help to improve the accuracy of nowcasts, especially if the resolution of the latent state is to be further increased.

This leads to yet another direction of future work: increasing the resolution beyond US states. At least two sources (GFT and TWTR) report at a sub-state level, and this data is left, for now, unused. Similarly, some sources are available at a higher temporal resolution than what I currently use. For example, GHT, TWTR, and CDCP are available daily; WIKI is available *hourly*. Although it may not be immediately obvious how to incorporate these data streams into the current framework, it is certainly desirable to produce nowcasts at as high of a resolution as possible.

Looking forward, the purpose of nowcasting is to provide timely and accurate situational awareness. While the implementation details are important, the way in which the results are used is equally important. As previously mentioned, these nowcasts are currently published in real-time through our API [151]. It is hoped that these nowcasts will be of use to the broader public, and to this end a web interface is needed. In the meantime, the nowcasts are fed as input

into the Epicast forecasting system—one of the main topics of Chapter 5.

# Chapter 5

# Forecasting Influenza Epidemics using Statistical Models and Human Judgment

> The reality in forecasting is that some of the time you are going to be wrong, and some of the time you are going to be badly wrong. You try to make it so that those bad forecasts are less and less frequent and that reliable, accurate forecasts become more the norm.
>
> *Jeffrey Shaman*

Much of this chapter is based on [167, 168].

## 5.1 Learning from the past

In temperate regions of the world, flu epidemics occur annually in the cold winter months. This is of great interest to public health, and in the US, CDC collects and publishes several indicators of flu activity every week, including in particular weighted percent influenza-like illness (wILI). Since 1997, CDC has collected and published wILI for the US as a whole and for the 10 HHS regions and 9 census regions within the US. The annual repeating pattern of flu epidemics becomes evident when plotting this data as a function of time (Figure 5.1).

Each epidemic is unique to some extent, but they all share some characteristic features. Consider the national timing (peak week) of each flu season. In each season (ignoring the 2009 pandemic) the peak week is observed between December and March. The maximum value of wILI within each season (the peak height) falls between 2% and 8%. There is usually one very strong peak, and possibly one or two smaller peaks. There is generally a large spike in wILI around the end of December and, to a lesser extent, around the end of November; these are the result of the "holiday effect", a measurement artifact. In several seasons, there is a small bump around March; this is usually due to a secondary outbreak of Influenza B. These are just a few examples—it is probably possible to find several more.

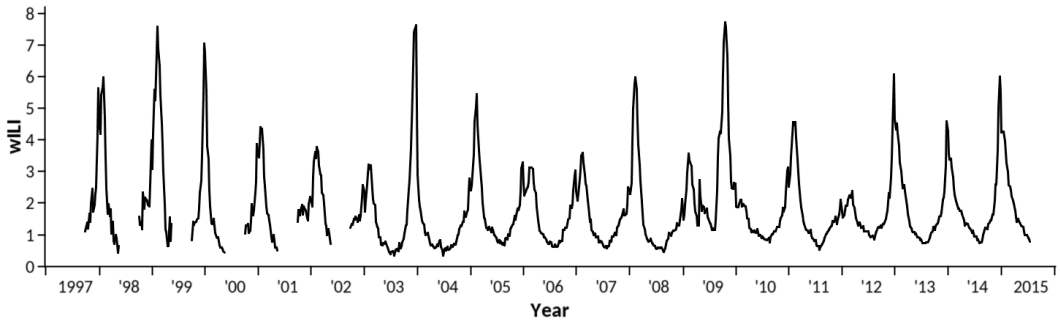


Figure 5.1: **Weekly time series of wILI in the US.** Data source: CDC/ILINet.

Given all of these commonalities, is it possible to make some kind of educated guess about the shape of the current epidemic? This is exactly the idea underlying the methods for epidemiological forecasting in this chapter.

## 5.2 An intuitive approach

### 5.2.1 Predictions, forecasts, and accuracy

There is an important distinction to be made between a *prediction* and a *forecast*. These words are often used interchangeably elsewhere, but in this thesis I use them to mean subtly different things. Of course, both make a statement about the future, but the difference has to do with (un)certainty. A prediction is a single point (technically, a single outcome), and it provides no indication of certainty. In the discrete case, it may assert that "wILI will peak on week 5", but it does not say anything at all about the possibility of wILI peaking on any other week. The more general continuous case is similar; a prediction makes an absolute statement about the future and says nothing about other potential outcomes. A forecast, on the other hand, is essentially the opposite; a forecast assigns a probability to all possible outcomes, but it makes no claim as to which particular outcome will take place. For example, a forecast might say "there is a 20% chance that wILI will peak on week *4*; 70% on week *5*; and 10% on any other week of the year".

This raises a question: which is better? Continuing with the example from above, suppose it eventually becomes known that wILI peaked on week *5*. The prediction is certainly correct, but is the forecast correct? Suppose instead that wILI peaked on week *4*. Now the prediction is definitely wrong, but what about the forecast? While a prediction can be correct or incorrect, a forecast cannot be said to be correct or incorrect, but only more or less accurate. Clearly, there is a need for a well-defined measure of accuracy.

For numerical predictions, one such measure of accuracy is the absolute size of the error. When there are many such predictions (for example, predicting wILI on the next $N$ weeks), mean absolute error (MAE) tells, on average, how far the prediction was from the truth. Another measure is squared error. Compared to absolute error which treats all errors equally, squared error penalizes large errors much more strongly than small errors. With many predictions, root

mean squared error (RMSE) gives an indication of the overall error. Both of these values are measured are in units of the original prediction; if wILI is being predicted, then MAE and RMSE will both be measured in units of wILI. Since these are both direct measurements of error, a value of 0 represents a perfect prediction. Given true outcomes $y$ and predictions $\hat{y}$, MAE and RMSE can be written as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|\,,$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}.$$

The gold standard for assessing forecast accuracy is the log score. A forecast consists of a set of mutually exclusive and jointly comprehensive outcomes and their associated probabilities. The log score is defined as the logarithm of the likelihood assigned to the outcome that actually happened. When there are many forecasts available, the mean log likelihood (MLL) is used to provide a combined likelihood score. Unlike MAE and RMSE, MLL is not measured in the units of the target being forecasted. Instead, it is measured in units that depend on the base of the logarithm used, which in the case of the natural logarithm is *nats*—natural units of information. Because the logarithm gives a negative number for values less than 1 (which is true for probabilities), log score is a negative value. Negating log score gives a positive number known as *surprisal* [169]. With either measure, a perfect forecast is given a score of 0. Given true outcomes $y$ and a set of forecasts $f$, MLL can be written as:

$$\text{MLL} = \frac{1}{N}\sum_{i=1}^{N}\log \Pr(y_i \mid f_i).$$

So to the original question: which is better? Predictions and forecasts provide different statements about the future, both of which are useful. A forecast, however, has more information than a prediction, and it is easy to derive a reasonable prediction from a forecast. For example, a prediction could be the forecasted outcome with highest probability—in other words, the distribution's mode. It is impossible, however, to produce a forecast given only a prediction. Consider also that a forecast is not limited to a unimodal distribution; it can be any shape with any number of modes. A prediction, however, is unable to capture anything other than a single point. For these reasons, it is preferable to have both, if possible; otherwise a forecast is preferred. Unfortunately, as I show below, it is often much more difficult to generate a full forecast than it is to predict a single value. Many strategies begin with a method of generating predictions and are then expanded to provide a forecast.

### 5.2.2 A first attempt

Suppose it is currently epiweek 2016w50. The time series of wILI up through epiweek 2016w49 is available (there is a 1–2 week lag in the reporting and publication process), and a prediction of wILI is requested for each week through the rest of the season—up to week 20 of the following

year, 2017w20. One good place to start is to find which past seasons appear to be most similar to the current season. In other words, find the seasons where wILI on epiweeks, say, w40–w49 approximately match the current season's wILI on the same epiweeks. Once these have been identified, a reasonable prediction could be this: wILI on epiweeks 2016w50–2017w20 will be the average wILI of the selected past seasons on the same weeks. The essential hypothesis here is that the future will resemble—at least to some extent—the past.

This simple idea underlies a diverse set of forecasting methodologies, including the *Pinned Spline* (SP), *Empirical Bayes* (EB), and *Epicast* (EC) methods. A brief overview of the spline method is given in Appendix A; I describe the latter two below.

## 5.3 Frameworks for epidemiological forecasting

### 5.3.1 An empirical Bayes approach

> The initial empirical Bayes idea was due to Ryan Tibshirani and Roni Rosenfeld; Logan Brooks, David Farrow (myself), and Sangwon Hyun contributed significantly to the implementation of this idea. The empirical Bayes method was one of our entries in CDC's 2013–2014 flu forecasting contest [86]—this is the version described in depth in [167] and which I describe more succinctly below. The empirical Bayes method has since been improved in many ways, most of which are due to Ryan Tibshirani and Logan Brooks. These improved versions have been used as entries in CDC's 2014–2015 and 2015–2016 flu forecasting contests and in OSTP's dengue challenge.

The empirical Bayes (EB) method can be summarized with the following sequence of operations:

1. **Build a model of past seasons.**
   From each wILI trajectory of past seasons, create a smoothed trajectory.
2. **Form a prior distribution of trajectories.**
   Create versions of past smoothed trajectories that have been altered in various ways.
3. **Estimate wILI on past weeks of the current season.**
   Use a mixture of traditional and digital surveillance to build a partial trajectory.
4. **Build a posterior distribution of trajectories.**
   Weight samples from the prior based on similarity to current season's partial trajectory.
5. **Report prediction and forecast for each target.**
   Measure target values on posterior trajectories and estimate their distributions.

Because wILI is an indirect and inherently noisy measurement of true influenza activity, it is desirable to model instead the underlying (but latent) signal which wILI approximates. This is modeled in the EB method as a smooth curve through the observed wILI time series. There are several methods that can be used to produce a smooth version of these curves: moving averages, kernel smoothers, smoothing splines (as in Appendix A), and many others. Here, we decided to use a relatively recent method called trend filtering [170].

Trend filtering has a couple of properties that make it desirable for smoothing wILI trajectories. First, it automatically, and more importantly *adaptively*, determines the amount of smoothing to apply over the time series. This is especially helpful because wILI trajectories before and after the epidemic are already relatively smooth, and a constant smoother may over-smooth these areas. On the other hand, the noise during the epidemic, and especially around the peak of the epidemic, is much higher, and a constant strength smoother is prone to under-smoothing this portion of the trajectory. Second, most smoothing methods have the undesirable side effect of pulling the peak wILI value down. (They also pull the minimum wILI value up, but this is less concerning since it is modeling the peak—not the trough—that is of greatest interest.) Trend filtering is capable of preserving the extreme values of the trajectory, which is critically needed since true wILI could be higher—not just lower—than observed wILI on all weeks, including the peak week.

The output from the smoothing process is twofold; in addition to smoothed trajectories, it also produces an estimate of the magnitude (standard deviation) of the noise, $\tau$. Therefore, after this step we end up with a set of smooth curves and their associated noise terms, which together form the basis of our prior (Figure 5.2).

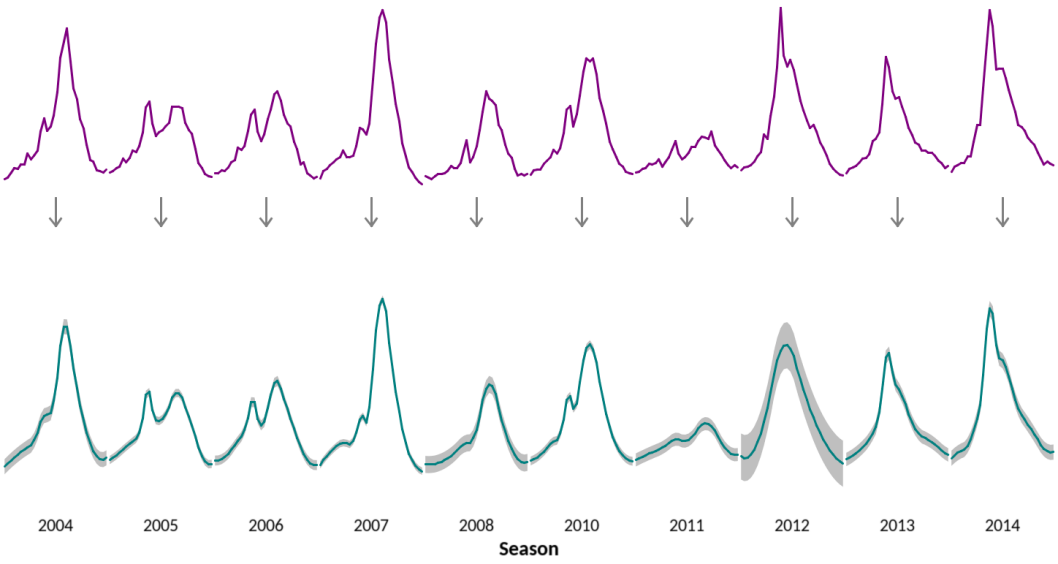


Figure 5.2: **Trendfiltered trajectory of selected seasons.** Trendfiltering gives a smooth model of latent influenza-like illness and an estimate of the magnitude of measurement noise.

Having formed a model of the trajectory of past seasons, the next step is to build a prior representing the space of all possible wILI trajectories. The key assumption here is that any wILI trajectory can be thought of as a variation on one of the past trajectories. These variations include five choices and transformations described below and illustrated in Figure 5.3.

1. **Starting shape.**
   Select uniformly at random the smooth trajectory of a past season.
2. **Peak magnitude.**
   Select uniformly at random the peak wILI value. The bounds of this range are defined by an unbiased estimate of a uniform distribution given empirical peak wILI values of past seasons. The trajectory is then scaled such that the original peak matches the selected peak. However, the scaling is *not* applied to wILI values below the region-specific baseline defined in advance by CDC.
3. **Shift in time.**
   Select uniformly at random the week on which wILI reaches its peak value. The bounds of this range are defined by an unbiased estimate of a uniform distribution given empirical peak weeks of past seasons. The trajectory is then shifted such that the observed peak week falls on the selected week.
4. **Epidemic pace.**
   Select uniformly at random a scale parameter in the time dimension from 75% to 125% speed. A slower pace corresponds to a longer epidemic duration, and a faster pace corresponds to a shorter epidemic duration. The trajectory is then scaled by the selected value, centered at the peak week—this operation does not change the peak week or height.
5. **Inject noise.**
   Select uniformly at random the noise magnitude of a past season, and add noise ($\sim \mathcal{N}(0, \tau)$) to the current trajectory.

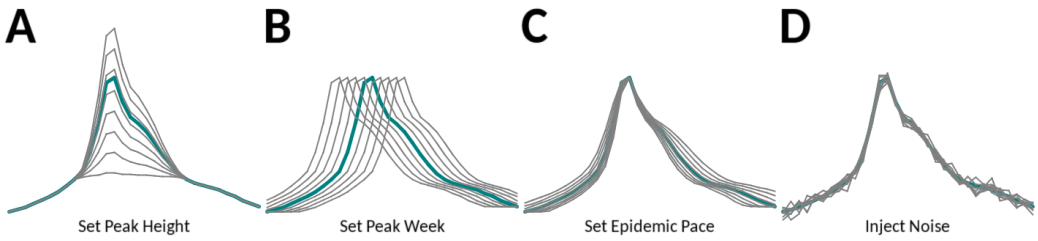


Figure 5.3: **A prior over wILI trajectories is built through a series of transformations of observed trajectories.** To begin, a random season is selected; this figure uses the 2013–2014 national season as an example (thick, aqua). Once a starting trajectory has been selected, it is altered (thin, gray) by **(A)** scaling peak height, **(B)** shifting peak week, **(C)** altering epidemic pace, and **(D)** injecting noise. For illustrative purposes, operations are shown independently; when generating a sample trajectory at run-time, operations are applied sequentially.

Everything up to this point—smoothing past trajectories and defining a prior distribution over trajectories—has been a function of the data that is available at the start of each flu season. It is therefore an off-line process that just needs to be performed a single time. These steps together make up the *empirical* portion of the EB method. The remaining tasks—estimating recent wILI,

building a posterior, and producing a forecast—are instead a function of the most recent data and are performed each time a new observation is made. For flu, using wILI provided by CDC, this is a process occurring each week.

The first of the on-line tasks is to estimate recent wILI. This estimate is necessary for a couple of reasons, the foremost among which is the 1–2 week lag between wILI data collection and reporting. Because of this lag, there could potentially be no official CDC estimate of wILI for the past several weeks. A further complication is that initial wILI reports are subject to change because only a subset of providers have reported to CDC the number of ILI cases they have observed. As more weeks pass, more reports are collected, and the value of wILI reported by CDC begins to stabilize. For this reason, the most recent 2–3 wILI values are somewhat suspect and are likely to be revised in coming weeks. Collectively, these issues were the primary impetus for the nowcasting framework in Chapter 4 which I developed later.

As discussed extensively in Chapter 4, digital surveillance can be used as a proxy for wILI, trading accuracy for timeliness. In our application of EB forecast the 2013–2014 flu season, we made use of Google Flu Trends (GFT) to estimate the five most recent wILI values. On any given forecasting week, CDC will have reported preliminary estimates of wILI for 3–4 of the past 5 weeks, and there will be no official estimates of wILI for the remaining most recent weeks. To estimate what final wILI will be on these most recent five weeks, I regressed GFT onto wILI on all prior weeks. I then used this model to compute an estimate of the most recent wILI values, given the five most recent GFT values. The end result is a hybrid trajectory of $n$ wILI values such that the first section ($n - 5$ weeks) is preliminary wILI as reported by CDC and the last section (5 weeks) is wILI as predicted by regression using GFT. For the 2014–2015 flu season, we did not incorporate GFT data.

The next on-line task is to build a posterior distribution of trajectories. This is achieved by sampling from the previously constructed prior and weighting samples by similarity to the fragment of the trajectory available for the current season. There are a number of metrics that can be used to define similarity in this context corresponding to associated assumptions about the distribution of the noise. Here we use the multivariate normal log likelihood with $\Sigma = \mathrm{I}$. Given, on week index $wk \in \{1,2,\cdots,52\}$, a sample from the prior *before injecting noise*, $p \in \mathbb{R}^{52}$, and the estimated wILI trajectory of the current season, $s \in \mathbb{R}^{wk}$, the similarity-based weight can be written as:

$$\text{weight}^{-1} \propto (p_{1..wk} - s)^T \Sigma^{-1} (p_{1..wk} - s),$$

$$\text{weight} \propto \left( \sum_{i=1}^{wk} (p_i - s_i)^2 \right)^{-1}.$$

Once a sample from the prior is assigned a weight, noise is injected as previously described. The noise is injected *after* computing similarity for the simple reason that dissimilar samples could, by chance, become superficially more similar with added noise.

The posterior distribution of weighted trajectories is built iteratively by repeatedly sampling and weighting trajectories from the prior. The application of Bayes' theorem in this step (defining a posterior in terms of a prior and a likelihood) is the source of *Bayes* in the EB moniker.

The required number of samples is not well-defined, and in practice we sample until target distributions converge. This generally requires on the order of $10^5$ iterations—about one minute of computation. Each of the prior transformations takes constant time, and the number of transformations computed is linear in the number of samples; similarly, each likelihood computation is linear in the number of observed weeks, and the number of likelihood computations is linear in the number of samples.

The final task is to output a point prediction and a forecast for each target. In the general case, we define the point prediction to be the weighted median of the target value measured on each posterior sample, corresponding to minimizing the expected absolute error of the prediction. We define the forecast as the smoothed distribution of weighted target values measured on all posterior samples. Reporting a distribution is a non-trivial task, and often a histogram is reported as a proxy. This has been the case, for example, in all flu contests to date. In the 2014–2015 flu contest, bins for timing targets were defined in one week intervals over the span of the flu season (with an additional bin representing no epidemic), and bins for wILI targets were defined in intervals of 1 wILI from 0–10 (with an additional bin representing wILI above 10). The details of the contest, definitions of the various targets, and analysis of EB forecasting accuracy are all discussed in the following sections.

### 5.3.2 A human judgment approach

> It is difficult to say who first articulated idea of soliciting epidemiological predictions from humans, but Ryan Tibshirani and Roni Rosenfeld both suggested use of pencil-and-paper surveys. Drawing inspiration from this idea, I conceived, designed, and implemented—and continue to run—the website and assimilation methodology known as "Epicast". The Epicast framework was one of the two winning entries in CDC's 2014–2015 flu forecasting contest [88], and it is currently one of our entries for CDC's 2015–2016 flu forecasting contest. The Epicast framework, with minimal modification, was also an entry in DARPA's 2014–2015 chikungunya prediction challenge.

Given the partially observed wILI trajectory of the current season, would a person be able to reasonably predict wILI through the remainder of the season? If many people were to make such predictions, would the aggregate forecast be accurate? How would such a forecast compare to forecasts produced by data-driven methods like Pinned Spline and Empirical Bayes? These questions inspired me to create *Epicast*, a website for collecting manual human predictions together with a methodology for producing an aggregate forecast using these predictions.

The Epicast website is designed to collect epidemiological predictions from a large set of volunteers. In the interest of collecting reasonable predictions from an informed crowd, Epicast includes many informational resources. Some of these are links to external resources: the CDC flu portal, the Wikipedia article on Influenza, data from ILINet, and a list of research articles on the topics of epidemiological forecasting. Another resource is an embedded Google News box on the user's home page which shows up to date popular media articles on the topic of "flu". Finally, the wILI trajectory of all past seasons is shown on the actual prediction interface (Figure 5.4).

Submitting personal predictions through the Epicast interface is intended to be easy, fast, and

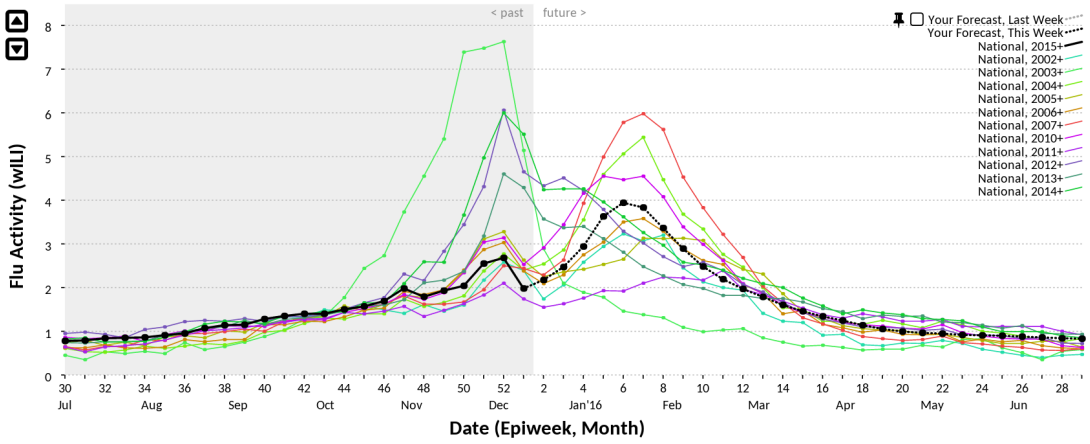


Figure 5.4: **A screenshot of the Epicast user interface.** On any given week, the past wILI trajectory of the current season is displayed (solid black). A user's prediction is a continuation of this trajectory over weeks in the future (dashed black). wILI trajectories of past seasons show the typical course of influenza epidemics (colored lines).

fun. Participants need only to click and drag to indicate what they think the wILI trajectory will look like on future weeks. Once a prediction has been made, the press of a single button will save the prediction and take the user to the prediction screen for the next region. Once predictions for all HHS regions and for the US as a whole ("national") have been collected, participants are taken back to their home page where they are told that they did a great job, thanked for their participation, and encouraged to share Epicast via email or social media. Each Friday, when CDC publishes new wILI values, we score user predictions and populate two leaderboards—the most accurate users, overall and on the last week—to encourage friendly competition and to motivate further participation.

The input that I collect from participating users is a set of wILI trajectories: a set of predicted wILI values on future weeks, separately for each region. The goal of the Epicast methodology is to aggregate these predictions to produce a probabilistic forecast. The way this is done is very similar to the way forecasts are generated by the Empirical Bayes method. The Epicast point prediction for any target is defined as the median of the target values measured on user predictions. The Epicast forecast for any target is a Student's t distribution with location equal to the median value (the point prediction), scale equal to the sample standard deviation of values, and degrees of freedom equal to the number of participants.

As with EB, the output of the Epicast system is twofold: a point prediction and a forecast over each target. Because the output of these systems is the same, it is possible to directly compare the forecasting accuracy of the two methods.

## 5.4 Results in forecasting the 2014–2015 flu season

### 5.4.1 Objectives, Targets, and Accuracy

As discussed in Chapter 2, CDC is currently hosting the third annual flu contest for forecasting wILI in the US. Our group has participated in all of these contests, and for the 2014–2015 season we submitted forecasts generated by the three previously mentioned forecasting methods: Pinned Spline (SP), Empirical Bayes (EB), and Epicast (EC). For the purposes of the contest, CDC required both a point prediction and a probabilistic forecast for a set of seven targets, separately for each of the ten HHS regions and for the US as a whole (eleven total regions). These targets were:

- **Onset Week.**
  The MMWR week on which wILI first reaches the epidemic threshold. This threshold, called a "baseline", is defined separately for each US region and is redefined each flu season. Technically, Onset Week is the first of three consecutive weeks on which wILI, rounded to one decimal place, is at or above the baseline in a given region.
- **Peak Week.**
  The MMWR week on which wILI reaches its maximum value.
- **Peak Height.**
  The value of maximum value of wILI. In other words, the height of the wILI trajectory on the Peak Week.
- **Lookaheads (4).**
  The next four wILI values. In other words, the height of the wILI trajectory at 1, 2, 3, and 4 weeks following the most recently reported wILI. Because of the lag in wILI reporting, the 1 Week Lookahead is technically a nowcast. I take advantage of this in Chapter 4.

Now that the contest is over, it has been revealed that Epicast was—at least by the metrics used by CDC—the winning system, not just among these three systems but among all seven competing entries in the flu contest. More precisely, Epicast was ranked highest in the four short-term targets, second-highest in the three season-wide targets, and achieved the highest combined score of any system. In what follows, I assess by a variety of metrics the forecasting performance of our three systems and show where each system excels and where each lags.

Each week during the 2014–2015 flu season I asked Epicast participants to predict wILI for each remaining week of the season. Each individually submitted prediction was a trajectory of varying length (depending on the week of submission) of wILI values, and I asked users to provide such predictions for each of the eleven total regions (treating US national as an additional region). From these submissions I produced an aggregate forecast over all seven targets as described above. In parallel, we ran the SP and EB systems on the same weeks to produce equivalent forecasts. Each Friday (usually), CDC published wILI for the preceding week, and on the following Monday we submitted forecasts separately for each of our three systems.

As previously discussed, I assess the quality of predictions in terms of mean absolute error (MAE). I assess the quality of forecasts in terms of (negated) mean log likelihood (MLL). To avoid unfairly penalizing the (at the time) surprising effects of backfill, I use not only the probability in the bin containing the true outcome, but also the probability assigned to one or two

adjacent bins. Although this definition of log score is somewhat unorthodox, it is at least consistent with the scoring rule adopted by CDC for the 2015–2016 flu contest. In the case of Onset Week and Peak Week, I consider the log score of the range of the actual Peak Week plus or minus one week (for example, if the Peak Week was 5, I compute the log likelihood of the probability assigned to a peak being on week 4, 5, or 6). Suppose that $\text{PkWk}_r^{\text{obs}}$ denotes the observed value of Peak Week in region $r$ and that $\text{P}(\cdots)$ represents the probability assigned by the forecaster to a given outcome. Then the score across all regions can be written as:

$$\text{score} = -\frac{1}{11}\sum_{r=1}^{11} \log \text{P}(\text{PkWk}_r \in [\text{PkWk}_r^{\text{obs}} - 1, \text{PkWk}_r^{\text{obs}} + 1]).$$

For the five wILI targets, I only have available a set of probability bins each of width 1 wILI, as this is what was required by CDC's flu contest. To determine which bins to include in the likelihood calculation, I select **a)** the wILI bin containing the actual value and **b)** the adjacent wILI bin nearest to the actual value. For example, the actual Peak Height in the U.S. National region was 6.002, and I select the two bins which together give the probability assigned to the event that actual Peak Height falls between 5 and 7. Suppose a forecast was made that $\text{P}(5 \leq \text{wILI} < 6) = 0.215$ and $\text{P}(6 \leq \text{wILI} < 7) = 0.412$; the log score assigned to this forecast is $-\log(0.215 + 0.412) = 0.467$. For Peak Height (and similarly for the Lookahead targets) across all regions:

$$\text{score} = -\frac{1}{11}\sum_{r=1}^{11} \log \text{P}(\text{PkHt}_r \in [\text{round}(\text{PkHt}_r^{\text{obs}}) - 1, \text{round}(\text{PkHt}_r^{\text{obs}}) + 1]).$$

While I strive to give equal treatment to each of the seven forecasting targets, Onset Week is somewhat problematic, and I handle this target separately for a couple of reasons. First, the epidemic onset occurred shortly after the start of the flu contest, and therefore the number of weeks on which we made predictions *before* the epidemic onset was much smaller than the number of weeks we made predictions ahead of, say, the epidemic peak. Second, the target of Onset Week is highly sensitive to ILINet backfill (recall Chapter 4 and Figure 4.1).

To further illustrate how backfill influences Onset Week, consider how changes to reported values of wILI cause the ground truth value of each target to change. The five wILI targets are somewhat robust to small changes in reported values; a forecast of 2.0 wILI is not so far off from a forecast of 2.1 wILI. By contrast, the target of Onset Week (and perhaps to a smaller extent, Peak Week) is fragile; small updates to published wILI values can have a large impact on the target week. For example, consider the scenario in HHS Region 1 (baseline = 1.2). On 2015w16 (20 weeks after actual onset), onset week measured on the most up to date data was 2014w48. One week later, an adjustment due to backfill caused wILI on 2014w49 to fall below the baseline, resulting in a new onset week of 2014w50. This small and delayed wILI revision, from 1.27 to 1.10, caused a two week shift in onset week (Figure 5.5). A similar situation happened in HHS Region 2. I return to the analysis of Onset Week after first comparing accuracy on each of the other forecasting targets.

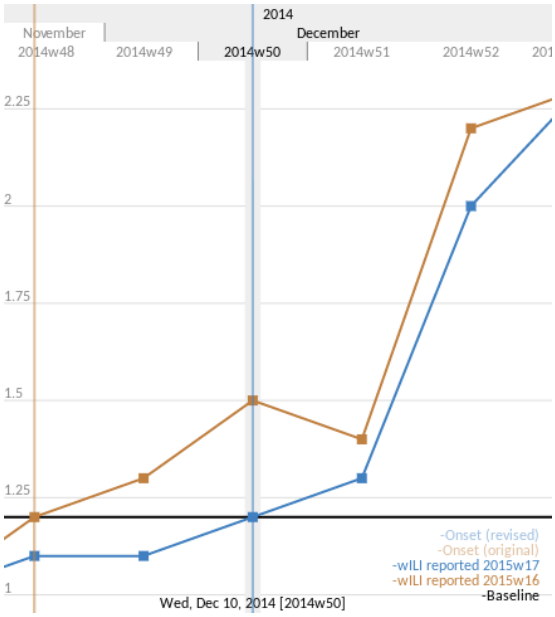


Figure 5.5: **Backfill in HHS region 1 causes a two week shift in Onset Week.** A large change (2 weeks) in onset week was caused by a small change (-0.17) in wILI.

### 5.4.2 Epicast Participation and Standalone Accuracy

In total, Epicast received 5,487 trajectories from a set of 48 volunteer participants during the 32 week period beginning on epiweek 2014w41 and ending on 2015w19. Participants varied in skill, from (self-identified) experts in public health, epidemiology, and/or statistics, to laypersons. Participation varied over time with an average of 16.1 participants per week (Figure 5.6). In the current analysis I did not handle expert and non-expert predictions differently, but I compare the performance of the two groups in a following section—the experts on average made slightly more accurate predictions.

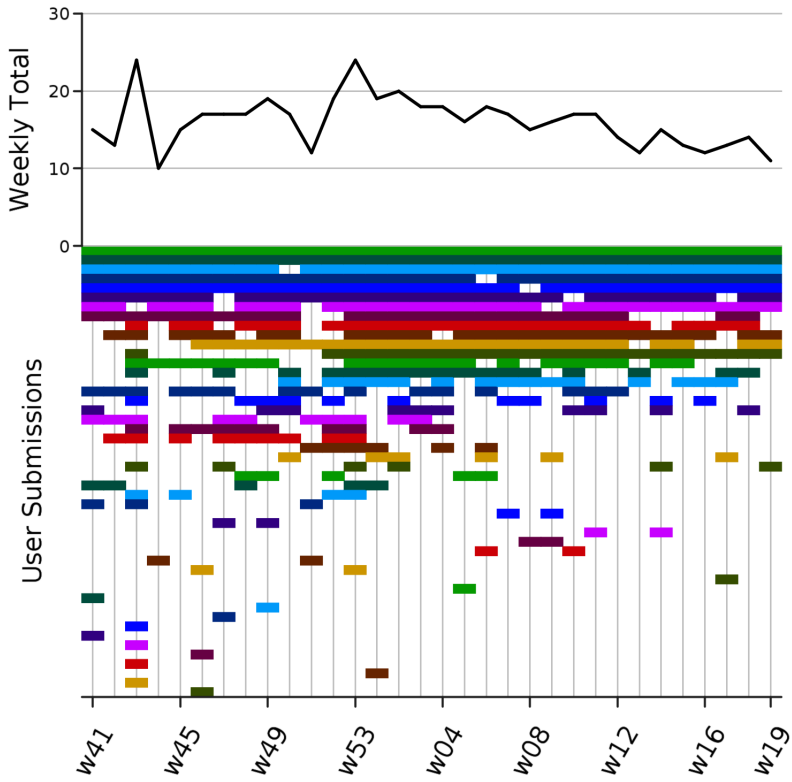


Figure 5.6: **Overview of 2014–2015 Epicast participation. Top:** total number of participants per week (avg: 16.1). **Bottom:** weekly submissions broken down by individual participants. Each participant is assigned a unique color which is used in subsequent figures.

To build an intuition for the standalone accuracy of the Epicast system, I test whether predictions fall within some range of the truth for each target. For the four short-term Lookahead targets, I count the fraction of the time that the predicted value falls within each range, grouped over all regions and weeks (Figure 5.7). The prediction is within 10% of the actual value just under half the time when predicting one week into the future; this falls to roughly one third of the time when predicting 4 weeks into the future. The trend is similar, though not as abrupt, at other accuracy thresholds. Accuracy within 50% is achieved near or above 95% of the time, even predicting up to 4 weeks ahead.

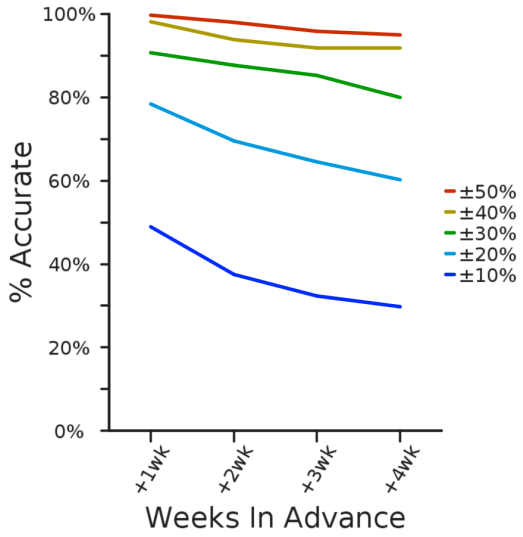


Figure 5.7: **Overall accuracy of Epicast for short-term targets.** The percent of regions and submission weeks ($n = 352$) where the Epicast point prediction was accurate within some range of the actual value is plotted as a function of short-term target.

To illustrate the varying difficulty of predicting each target throughout the season, I next consider a similar measure of accuracy as a function of *lead time*—the number of weeks preceding the Peak Week within each region (Figure 5.8). For 2, 3, and 4 weeks ahead, the lead time with lowest accuracy is 2, 3, and 4 weeks before the Peak Week, respectively, which suggests that there is a distinct challenge in forecasting the Peak Height. All short-term targets appear to be more accurate early and late in the season and less accurate around the Peak Week; this is to be expected, because there is significantly more volatility around the peak of the epidemic. The situation is quite different for the season-wide targets in which accuracy approaches 100% within two weeks after the peak. Accuracy of Peak Height prediction is initially low, but rapidly increases starting around 5 weeks before the peak. I defined accuracy in Peak Week slightly differently; it is the fraction of the regions in which the predicted Peak Week was within $N$ weeks of the actual Peak Week ($N \in \{1, 2, 3, 4, 5\}$). The situation for Peak Week closely matches that for Peak Height, and again I find that accuracy rapidly increases starting around 5 weeks before the peak and reaches its maximum two weeks after the peak.

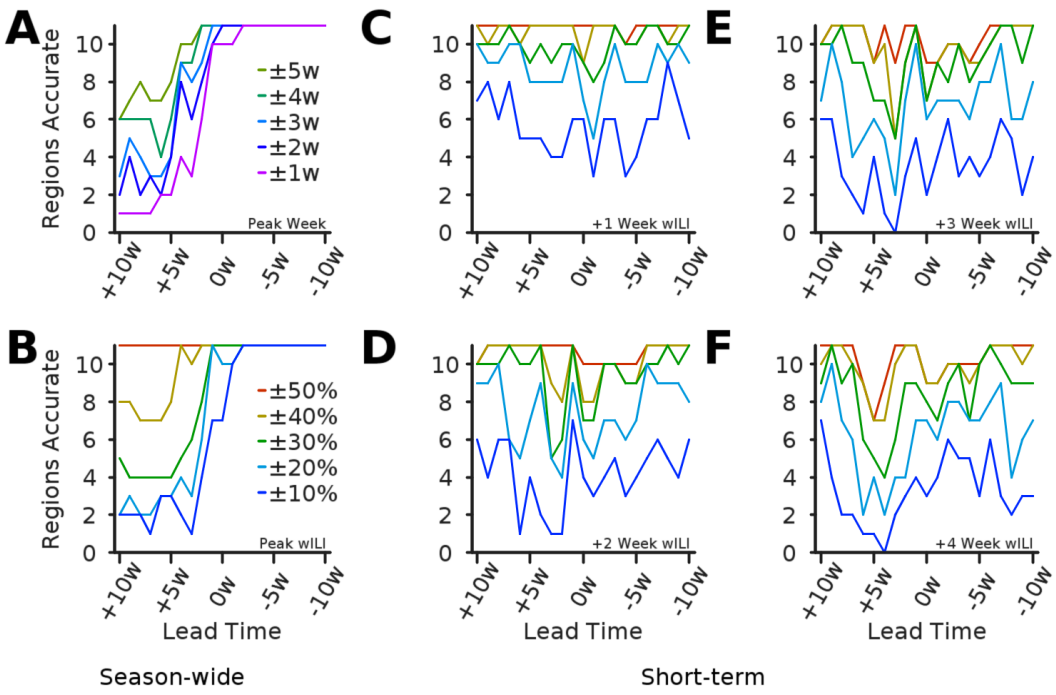


Figure 5.8: **Accuracy of Epicast by lead time for all targets.** Accuracy, as the number of regions where the Epicast point prediction was accurate within some range of the actual value, is plotted as a function of lead time. For timing targets, the range is the actual value plus or minus 1, 2, 3, 4, or 5 weeks; for wILI targets, the range is 10%, 20%, 30%, 40%, or 50% above and below the actual value. Subplots show accuracy in **(A)** Peak Week, **(B)** Peak Height, and **(C, D, E, and F)** wILI at 1, 2, 3, and 4 Week Lookaheads, respectively.

To more quantitatively evaluate the Epicast method, I calculate separately for each target Epicast's MAE across regions as a function of lead time (Figure 5.9). In agreement with previous results, MAE in season-wide targets generally decreases with lead time and is highest in short-term targets when predicting the peak. Additionally, MAE is elevated on the Peak Week (lead time = 0) across all short-term targets, indicating a relative increase in uncertainty immediately after the true peak (which is not known at the time to be so).

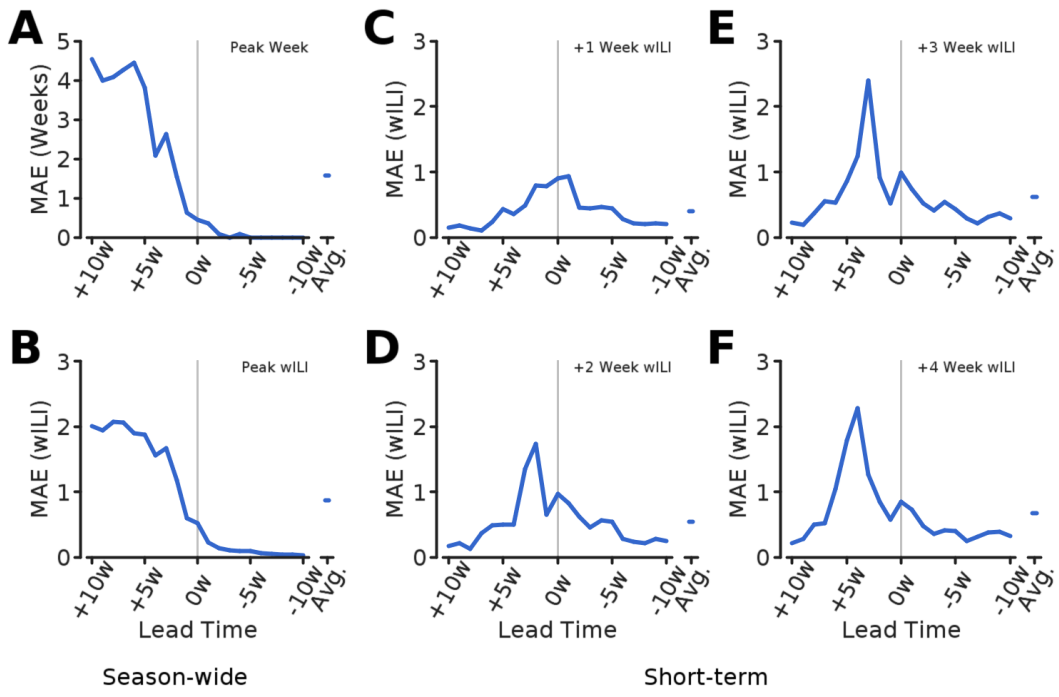


Figure 5.9: **Mean absolute error of Epicast by lead time for all targets.** Mean absolute error across regions ($n = 11$) is plotted as a function of lead time. Subplots show MAE in **(A)** Peak Week, **(B)** Peak Height, and **(C, D, E, and F)** wILI at 1, 2, 3, and 4 Week Lookaheads, respectively.

### 5.4.3 Comparison of Accuracy Between Forecasting Methods

To contextualize the accuracy of the Epicast method, I compare Epicast accuracy with the accuracy of individual participants and against the accuracy of two statistical forecasting methods. The statistical systems are the previously mentioned Pinned Spline (SP) and Empirical Bayes (EB) methods. The main challenge in presenting these results is that the space in which comparisons can be made consists of several orthogonal dimensions: regions (national + 10 HHS regions), targets (Peak Week, Peak wILI, and wILI 1–4 weeks ahead), submission weeks (depending on target, up to 32), and metrics (MAE and log score).

Concisely representing system performance requires the non-trivial task of reducing this dimensionality, otherwise it would require thousands of separate figures of merit. Several confounding issues impede aggregation along any one axis: forecasting difficulty varies over time as the season progresses, the various regions may peak at different times in the season, long-term targets are inherently more difficult to predict than short-term targets, and targets are measured in different units.

To work around these complications in the case of point predictions, I rank systems and participants in terms of absolute error and perform subsequent analysis on the relative ranking assigned to each forecaster. More specifically, I consider the pairwise ranking in absolute error

of Epicast versus individual participants or statistical methods. For each lead time, region, and target, I ask whether Epicast or the competitor had a smaller absolute error, and I measure the fraction of instances where Epicast had the smaller error—a "Win Rate".

To assess the statistical significance of each result, I use a Sign test with the null hypothesis that the pair of forecasters is equally likely to win (having smaller error). It should be noted that this test assumes that all observations are independent, but results across adjacent weeks, for example, are likely to be correlated to some extent. Overall, considering all targets, Epicast has lower error than all individual participants and both statistical methods (Figure 5.10). A similar result holds when only considering the four short-term targets. In the two season-wide targets, Epicast does well overall, but a small set of participants are more accurate than Epicast (one significantly so). In all cases, Epicast outperforms—often significantly—the two statistical systems.

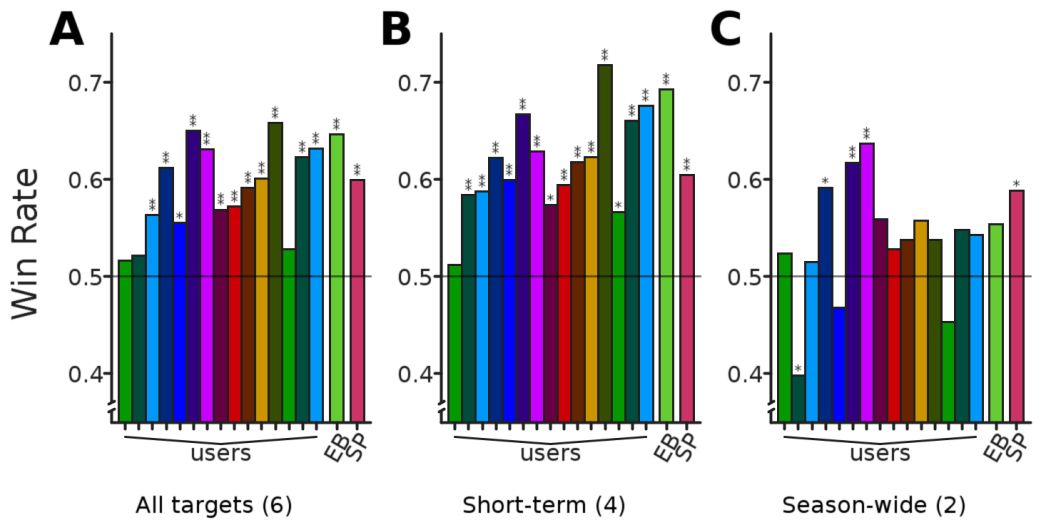


Figure 5.10: **Epicast Win Rate against individual human predictions and competing systems.** All plots show, for each predictor (users participating on at least half of the weeks, and two statistical systems), Win Rate: the fraction of instances where Epicast had lower absolute error than the competitor, across all regions and lead times ($n = 231$ per target). Statistical significance is determined by Sign test; *: $p < 10^{-2}$; **: $p < 10^{-5}$. Subplots show Win Rate considering **(A)** all targets, **(B)** the four short-term targets, and **(C)** the two season-wide targets.

Next, I compare forecasts in terms of log scores, however my analysis in this context is limited to only Epicast and Empirical Bayes as these are the only two systems for which we have reliable forecasts in addition to simpler point predictions. (The Pinned Spline system also produced forecasts, but these were often quite unreasonable.) I compute the average of these log scores for Epicast and EB for each target and for each value of lead time (Figure 5.11). To further contextualize the log score of each system, I show also the log score of a hypothetical system in which uniform probability is assigned to all plausible outcomes. For Peak Week, I define

this as a uniform distribution over weeks 2014w46 through 2015w12 ($p = \frac{1}{20}$ per week), and for the wILI targets I define this as a uniform distribution over wILI from 0 to 12 in increments of 1 ($p = \frac{1}{12}$ per bin). This uniform system provides a lower bound on the performance of a reasonable forecaster. Although Epicast has the smallest average log score for most targets, EB more consistently scores within the bound of the uniform system. EB particularly outperforms Epicast on season-wide targets during the weeks preceding the peak, but the trend reverses after the peak as Epicast more rapidly converges on the true value. On the short-term targets, Epicast generally scores better than EB, except when predicting wILI of the Peak Week.

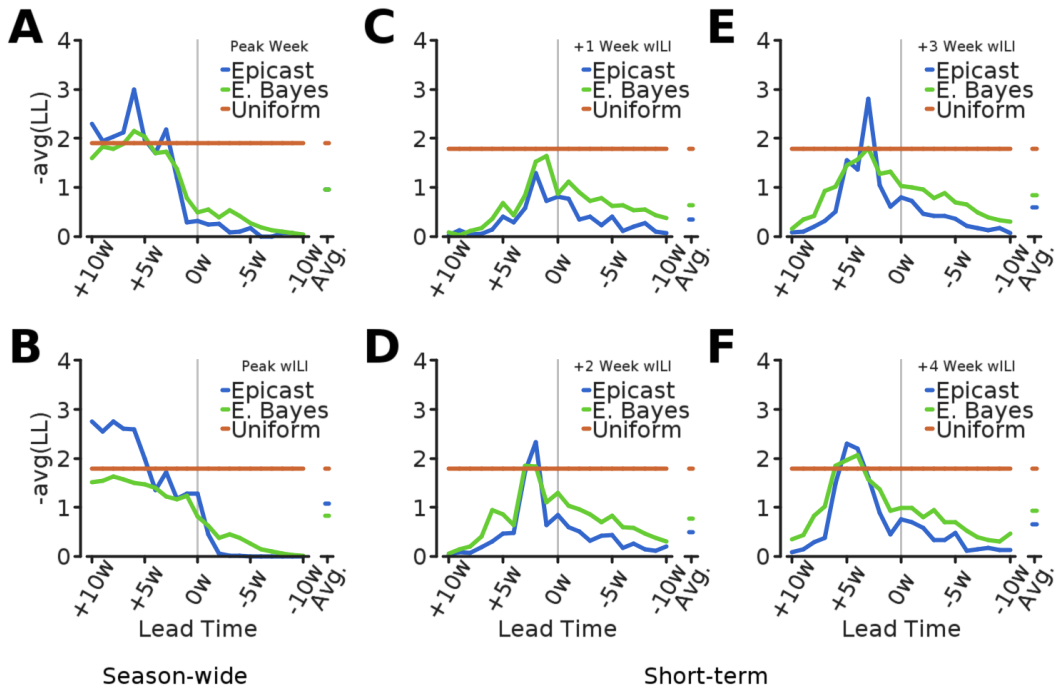


Figure 5.11: **Comparison of log scores for Epicast and Empirical Bayes.** Log scores, averaged across regions, of Epicast, Empirical Bayes, and the uniform forecast are plotted as a function of lead time. The average log score across time by system is shown to the right of each plot. Subplots show log scores by **(A)** Peak Week, **(B)** Peak Height, and **(C, D, E, and F)** wILI at 1, 2, 3, and 4 Week Lookaheads, respectively.

Finally, I summarize the performance of several forecasting strategies in terms of MAE in Table 5.1. I compare MAE separately for each forecasting target for each of Epicast, Pinned Spline, Empirical Bayes, and a simple "Baseline" approach.

The baseline prediction for any given target is computed as follows. Start with the wILI trajectory of each past season, starting in 2003 and excluding the 2009 pandemic. Each of these ten curves starts on week 30 of one year and ends on week 29 of the following year, truncated to 52 weeks on years with 53 weeks. At any point during the flu season, wILI has only been observed on some of the weeks. The baseline replaces wILI on all ten curves with the wILI

values that have been observered so far in the current season. This results in a set of ten composite curves where the "left" side is wILI observed in the current season and the "right" side is wILI reported in one of the past flu seasons. Finally, the target value of interest (say, Peak Week) is measured on each of the ten composite curves. The Baseline prediction is simply the median of the ten target values.

Despite the relative simplicity of the Baseline approach, on rare occasion it outperforms our more sophisticated systems. Overall, however, the baseline is generally inferior to the other approaches. Epicast shows particularly good performance, being ranked #1 in five of the seven targets and #2 in the other two targets. It appears that—at least for MAE in the 2014–2015 flu season—the general ordering of these systems from best to worst is: Epicast, Pinned Spline, Empirical Bayes, and Baseline.

| Target | Epicast | Spline | E. Bayes | Baseline |
|---|---|---|---|---|
| **+1 Week wILI** | $0.330^{1}$ | $0.380^{2}$ | $0.589^{3}$ | $0.795^{4}$ |
| **+2 Week wILI** | $0.439^{1}$ | $0.539^{2}$ | $0.700^{3}$ | $0.813^{4}$ |
| **+3 Week wILI** | $0.498^{1}$ | $0.650^{2}$ | $0.794^{3}$ | $0.833^{4}$ |
| **+4 Week wILI** | $0.557^{1}$ | $0.740^{2}$ | $0.879^{4}$ | $0.856^{3}$ |
| **Peak wILI** | $0.682^{2}$ | $0.992^{4}$ | $0.710^{3}$ | $0.641^{1}$ |
| **Peak Week** | $1.284^{2}$ | $1.244^{1}$ | $1.605^{4}$ | $1.491^{3}$ |
| **Onset Week** | $0.585^{1}$ | $0.722^{2}$ | $0.989^{3}$ | $0.991^{4}$ |

Table 5.1: **Mean absolute error by forecasting target and system.** For each of the seven forecasting targets (rows), the MAE of that target over all regions and submission weeks is shown for each system (columns). The rank of each system is indicated by a superscript number.

### 5.4.4 Results in forecasting Onset Week

Here I repeat the standalone and comparative analyses of Epicast performance in forecasting Onset Week (Figure 5.12). Due to an early epidemic onset in most regions, the analysis by lead time is limited to a window of six weeks. As before, I find that when considering pairwise ranking in absolute error, no individual user or system has a statistically significant Win Rate over Epicast. Two users do win over half the time, but the result does not reach the $p < 10^{-2}$ threshold of significance. On the other hand, Epicast beats the two statical systems—one significantly. As with the other season-wide targets, I find that MAE in Onset Week falls during the weeks preceding the onset and then levels off. Unlike the other targets, error in onset week, by all measures, never reaches zero; predicted onset is always off by more than one week in at least one region, as is particularly evident in the plot of average log score. The reason for this apparent shortcoming is, as discussed throughout, due to backfill.

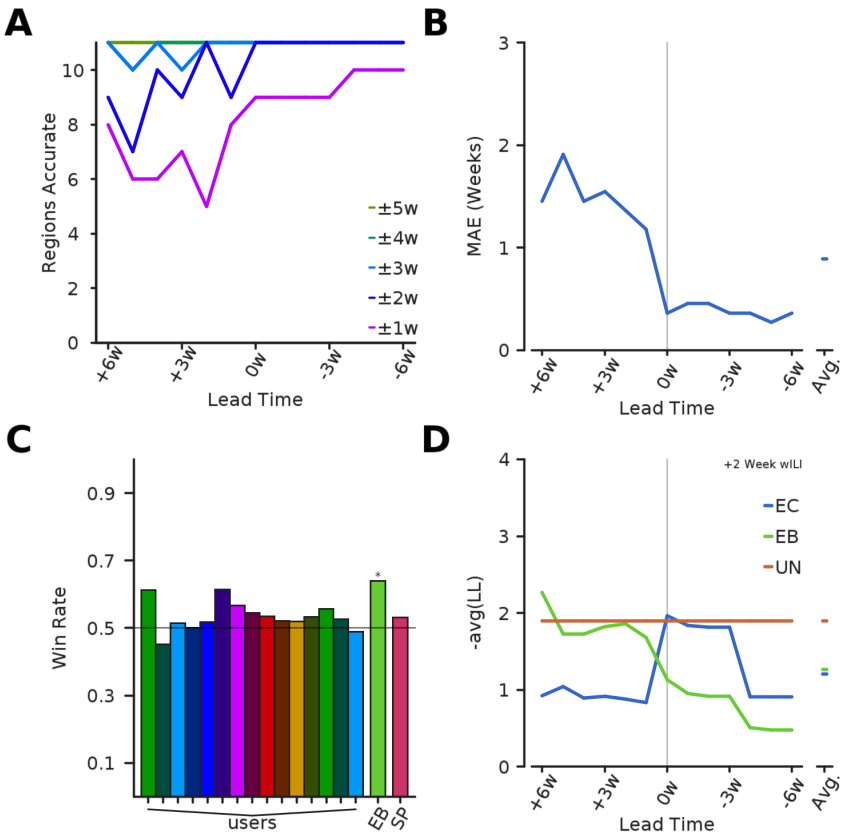


Figure 5.12: **Accuracy in forecasting Onset Week.** Epicast performance on an additional season-wide target, the week of epidemic onset. **(A)** As in Figure 5.8, the number of regions in which Epicast's prediction falls within a given range of the actual onset week, plotted as a function of lead time relative to epidemic onset. **(B)** As in Figure 5.9, MAE (across regions) in onset is plotted as a function of lead time relative to epidemic onset. **(C)** As in Figure 5.10, pairwise Win Rate is shown for Epicast against individual users and statistical systems. **(D)** As in Figure 5.11, average log score is shown for Epicast (EC), Empirical Bayes (EB), and the Uniform System (UN) as a function of lead time relative to epidemic onset.

## 5.5 Adaptive extensions to the Epicast framework

The surprising accuracy of the Epicast method raised several interesting questions. Does past experience or expertise influence prediction accuracy? Are some participants inherently better at predicting flu than others? Would it be better to trust the predictions of some participants more than predictions of others? I attempt to answer these questions in the following retrospective analysis.

### 5.5.1 The relative accuracy of expert and non-expert predictions

An interesting question that arises in the context of "Wisdom of Crowds" forecasting is whether "experts" (by some definition) make better forecasts than non-experts. Within the online Epicast user interface, I gave users the option to self-classify as having background or expertise in up to five areas: epidemiology, statistics and/or machine learning, virology, public health, and influenza. Of the 48 active users, 25 claimed expertise in at least one area, and the other 23 did not claim expertise in any area.

To assess the relative performance of the "experts" versus the "non-experts", I built forecasts using only predictions made from each group. I show MAE (Figure 5.13) and MLL (Figure 5.14) of each group as a function of lead time for each target. Perhaps unsurprisingly, the expert group generally has MAE and MLL less than or equal to that of the unmodified Epicast forecast and the Epicast based only on non-expert inputs. The most striking difference between the two groups appears when forecasting Peak Week, especially at a long lead time. On the other hand, accuracy between the two groups is essentially indistinguishable for the 1- and 2- Week Lookahead targets. The difference when predicting Peak Height and wILI at 3- and 4- Week Lookaheads is small, but noticeable.

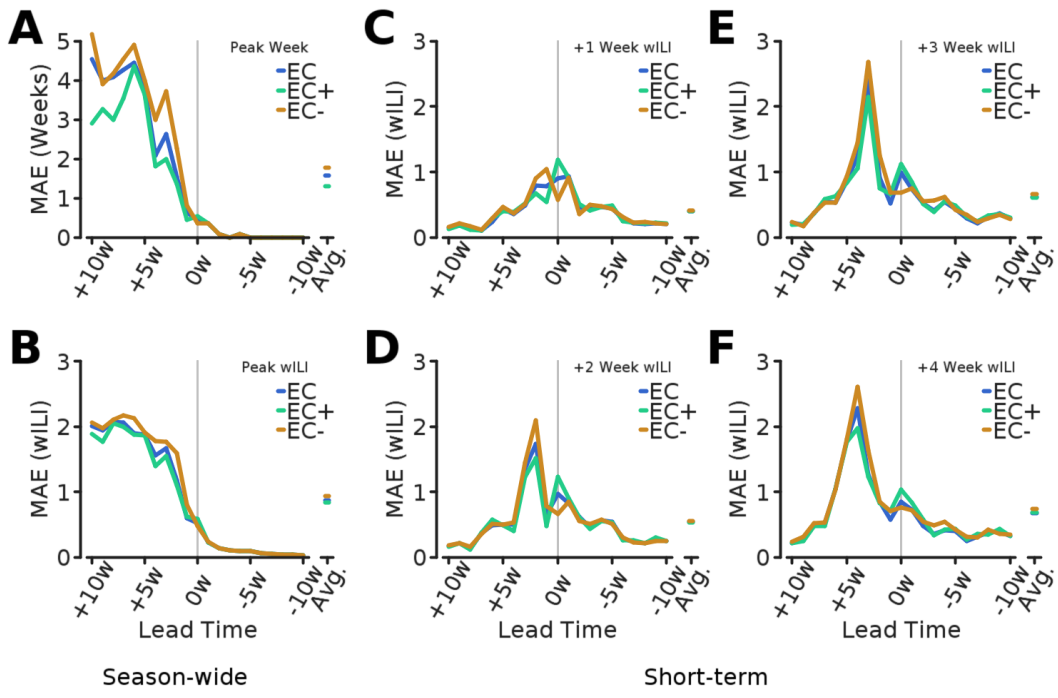


Figure 5.13: **Expert versus Non-expert MAE.** MAE is plotted for unmodified (all users) Epicast (EC), expert-based Epicast (EC+), and non-expert-based Epicast (EC-) as a function of lead time relative to the Peak Week. As in Figure 5.9, panels **A, B, C, D, E, and F** show MAE separately for each target.

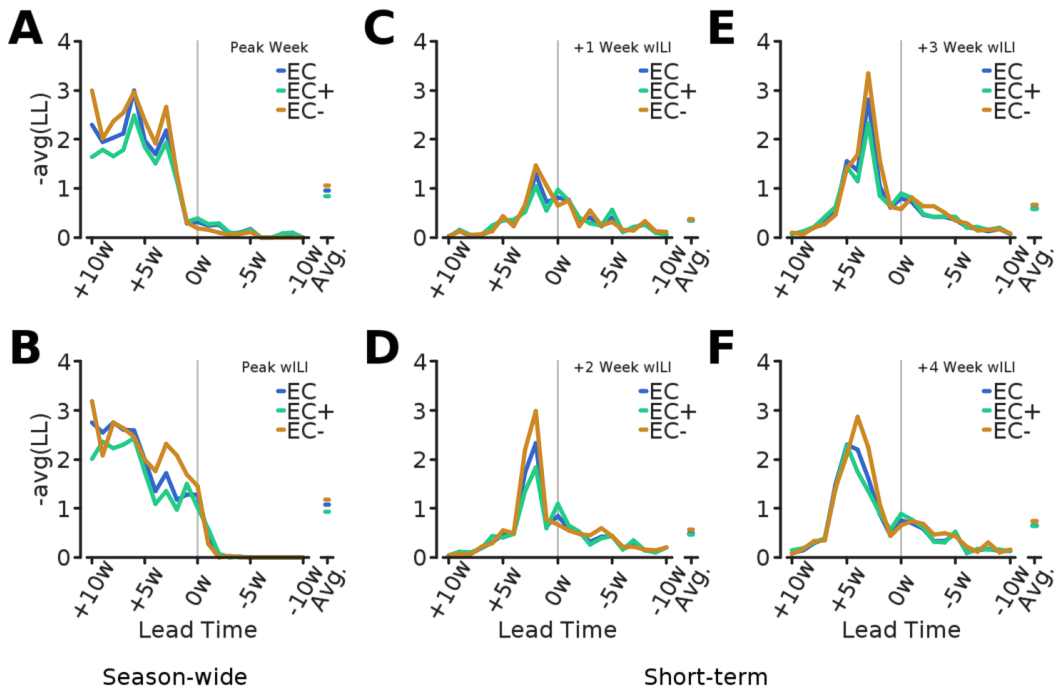


Figure 5.14: **Expert versus Non-expert MLL.** Mean log-likelihood is plotted for unmodified (all users) Epicast (EC), expert-based Epicast (EC+), and non-expert-based Epicast (EC-) as a function of lead time relative to the Peak Week. As in Figure 5.11, panels **A, B, C, D, E, and F** show log score separately for each target.

To determine whether these observations are statistically significant, I use a Sign test as before, this time separately for each target (Figure 5.15). Neither expert nor non-expert versions of Epicast have a significantly higher Win Rate than the Epicast built from all users. On other hand, none of the Win Rates reaches the $p < 10^{-2}$ threshold of significance. It is, however, clear that the expert group loses to unmodified Epicast less frequently than the non-expert group. This raises the question of whether some subset of—or a special weighting of—users could significantly improve Epicast performance. I consider this question in the next section.

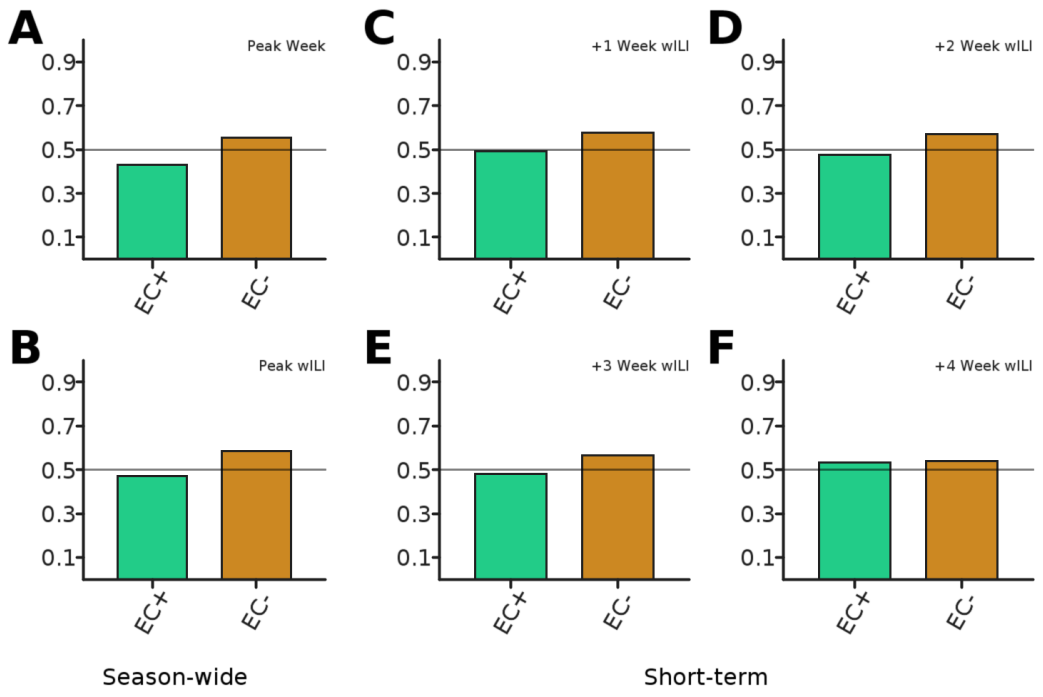


Figure 5.15: **Expert versus Non-expert Win Rate.** Win Rate of the unmodified (all users) Epicast system is plotted against expert-based Epicast (EC+) and non-expert-based Epicast (EC-). Panels **A, B, C, D, E, and F** show Win Rate separately for each target.

### 5.5.2 A weighting scheme based on expertise and past performance

Given that some users consistently outperform other users, is it possible to apply an adaptive weighting scheme to user predictions so that the overall accuracy of the Epicast is significantly improved? The primary obstacle to implementing such a scheme in *real-time* is that, due to backfill, it is not possible to know which users are outperforming their peers. Still, it is possible to estimate relative user performance using preliminary data and boost the weight of (hypothesized) high-accuracy users from week to week. I retrospectively attempt such a scheme below. Because I am limited to one season of data, I am unable to exhaustively test adaptive weighting schemes; however, I think the one described here, while *ad-hoc*, is a reasonable approach.

The original Epicast can be thought of as having a static weighting system wherein each user is given a weight of 1. Now I weight users based on two criteria: those who self-identify as experts, and those who had the lowest absolute error 2 weeks ago for a 3-week-ahead prediction of wILI (performance on the 3 Week Lookahead wILI target, after having seen preliminary "truth" values for two weeks). I select this particular target and timing for two reasons. First, it is difficult to differentiate user performance on very-short-term targets; predictions are very clustered at one week ahead and begin to spread out as the length of the prediction increases. At 3 weeks ahead it becomes reasonable straightforward to differentiate user performance. So

then, why use the 3 Week Lookahead target and not the 4 Week Lookahead target? This leads to the second reason: I need data that is as finalized as possible to determine who is performing well. Because of the previously discussed backfill issue, initial wILI estimates are subject to revision. Reliability increases (backfill adjustments decrease) over time. The initial report is too unreliable to determine who is doing well, so I wait one additional week for the wILI value to settle a little bit closer to its final value. Finally, there is the additional constraint that I want to know as soon as possible who is performing the best—I want to know as early as possible in the season who the accurate users are, and I want to be able to quickly adapt to changes in their performance. This method will allows for the ranking of users with a total lag of 5 weeks. To be more concrete, the proposed weighting scheme is this: $w_u = 1 + e_u + r_u$, where $w$ is the weight given to a prediction, $e$ is 1 for (self-identified) experts (0 otherwise), and $r$ is 1 for the top (arbitrarily) 5 users in terms of MAE on 3 Week Lookahead predictions (0 otherwise)—for each user $u$.

To illustrate the adaptive nature of the weighting scheme proposed above, I show a moving average (over 5 weeks) of user weight as a function of submission week for the 5 users with the highest number of submissions (Figure 5.16). As expected, some users are consistently given more weight than other users, and these weights change in response to accuracy of past predictions. To determine whether the Epicast built with these weights is better than the original Epicast, I show MAE (Figure 5.17) for each target as before. Unsurprisingly, the weighted Epicast outperforms the original, unweighted Epicast. However, it is difficult to say if there is any real advantage to using this particular weighting scheme, especially when considering the short-term targets. With the exception of the 4 Week Lookahead target, the weighted Epicast achieves a slightly higher Win Rate than the unweighted Epicast, but none of the Win Rates reach significance at the $p < 10^{-2}$ level. Still, these results suggest that there may be value in considering both user skill (self-classification as having expertise) and past performance (accuracy on previous predictions) when aggregating user predictions. I suspect that the advantage of a weighted Epicast would be much more meaningful with a larger, and more diverse, set of participants.

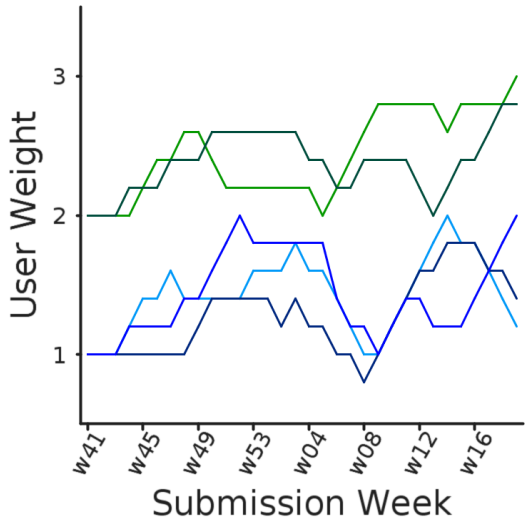


Figure 5.16: **User Weight over Time.** The five week moving average of user weight is shown for the 5 most active users.

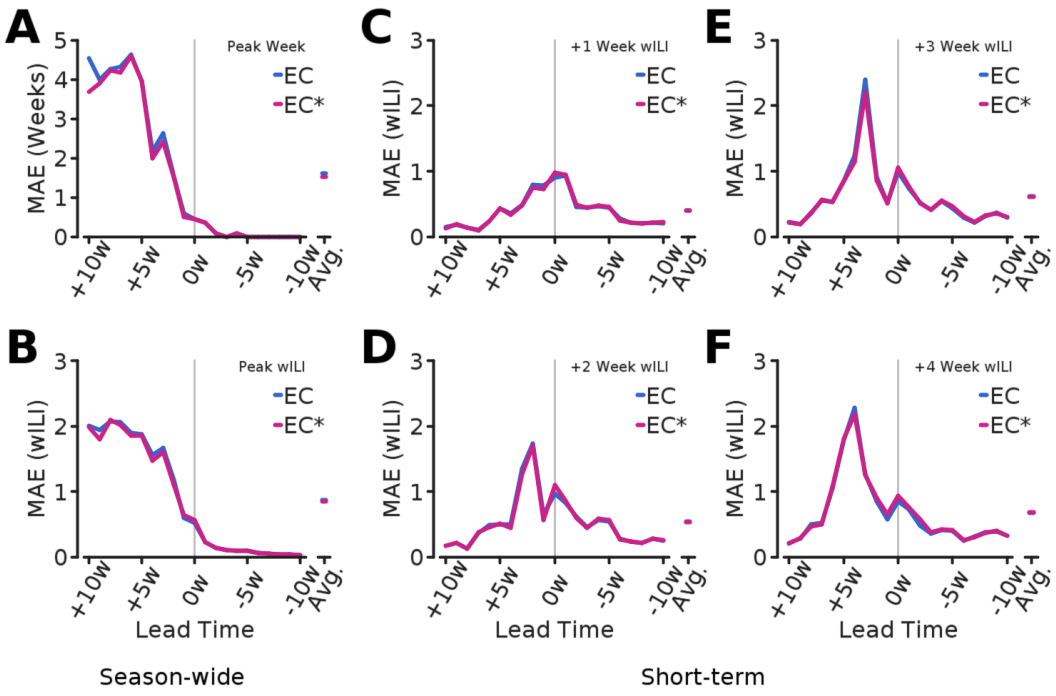


Figure 5.17: **MAE of Weighted Epicast.** MAE is plotted for unweighted Epicast (EC) and weighted Epicast (EC*) as a function of lead time relative to the Peak Week. As in Figure 5.9, panels **A, B, C, D, E, and F** show MAE separately for each target.

## 5.6 Epidemiological forecasting of other diseases

Although the epidemiological forecasting systems described in this chapter were designed to forecast influenza in the US, the methods are readily generalizable to other diseases and locations. To demonstrate this, I show the results of two additional forecasting challenges in which our group participated: the OSTP dengue challenge and the DARPA chikungunya challenge. For the former, we used an approach based on a hybrid methodology of Empirical Bayes and Pinned Spline; for the latter, I used a modified version of Epicast to collect predictions from a selected set of experts in fields relating to epidemiology and vector-borne diseases. Both of these challenges focused on predicting disease outbreaks in tropical America.

### 5.6.1 OSTP Dengue Challenge

> Our forecasting system for the dengue challenge was a mixture of two systems: Empirical Bayes and Pinned Spline. Logan Brooks managed the Empirical Bayes system and compiled and submitted forecasts, Ryan Tibshirani provided the original Pinned Spline implementation, and David Farrow (myself) adapted the Pinned Spline framework to handle external covariates for the dengue challenge.

The dengue challenge, unlike the influenza and chikungunya challenges, was retrospective; we were given training data up until 2009, and we were asked to retrospectively forecast, or "backcast", dengue incidence from 2009–2013. As with the flu contest, forecasts were required for multiple locations. These locations were Iquitos, Peru and San Juan, Puerto Rico. There were three targets for which we were asked to provide both point predictions and forecasts:

- **Peak Week.**
  The week of highest dengue incidence.
- **Peak Incidence.**
  The highest dengue incidence. In other words, the height of the dengue case trajectory on the Peak Week.
- **Total Cases.**
  The total number of cases (both confirmed and suspected) throughout the forecasting period.

The essence of our methodology was as previously described, with just a few exceptions. First, we combined the outputs of the two statistical systems. Second, we included covariates of temperature and precipitation in the Pinned Spline method.

The combining process took three inputs: forecasts generated by the Empirical Bayes system, forecasts generated by the Pinned Spline system, and an empirical prior over target values. This empirical prior is a type of baseline forecaster that does not rely on any observations within the current season; instead, it simply forecasts, for any given target, the empirical distribution of the target values which have been observed in past years. The point prediction of the empirical prior, like many of our forecasting systems, is just the median of the observed target values. Leave-one-out cross-validation was used to find the optimal mixture weights of the three systems. These weights (summing to 0.99) were computed separately for each evaluation metric (MAE and log-likelihood) and were recomputed for each forecasting week. The remaining 0.01 weight was

assigned to a uniform distribution over all outcomes so as to prevent assigning a probability of 0 to any event.

The climatic data consisted of daily temperature and precipitation (and many other) values for both locations. Before using the precipitation data as a covariate, I first applied a smoothing filter to model cumulative precipitation as a proxy for the amount of standing water. This operation, which produces new observations $z$ from original observations $y$ with tuning parameter $\alpha$(=$0.25$), can be expressed as:

$$z_t = y_t + \alpha z_{t-1}$$

I did this because **a)** the raw precipitation data was not strongly correlated with dengue incidence and **b)** dengue is spread by mosquitoes, which require standing water for breeding. The two covariates (temperature and standing water) were aggregated at a weekly timescale to match the incidence data, and they were used as covariates in the spline regression method. This deviates from the Pinned Spline system described in Appendix A in that the original Spline method only used a set of spline basis functions as covariates, but the Spline method for this challenge included these two additional climatic covariates.

After evaluation of all submissions, the challenge organizers shared with us our log scores (Figure 5.18). Our hybrid system performed well across all targets and locations, particularly in forecasting total incidence in San Juan. A more detailed comparison of forecasting accuracy across participating teams is expected in a forthcoming publication.

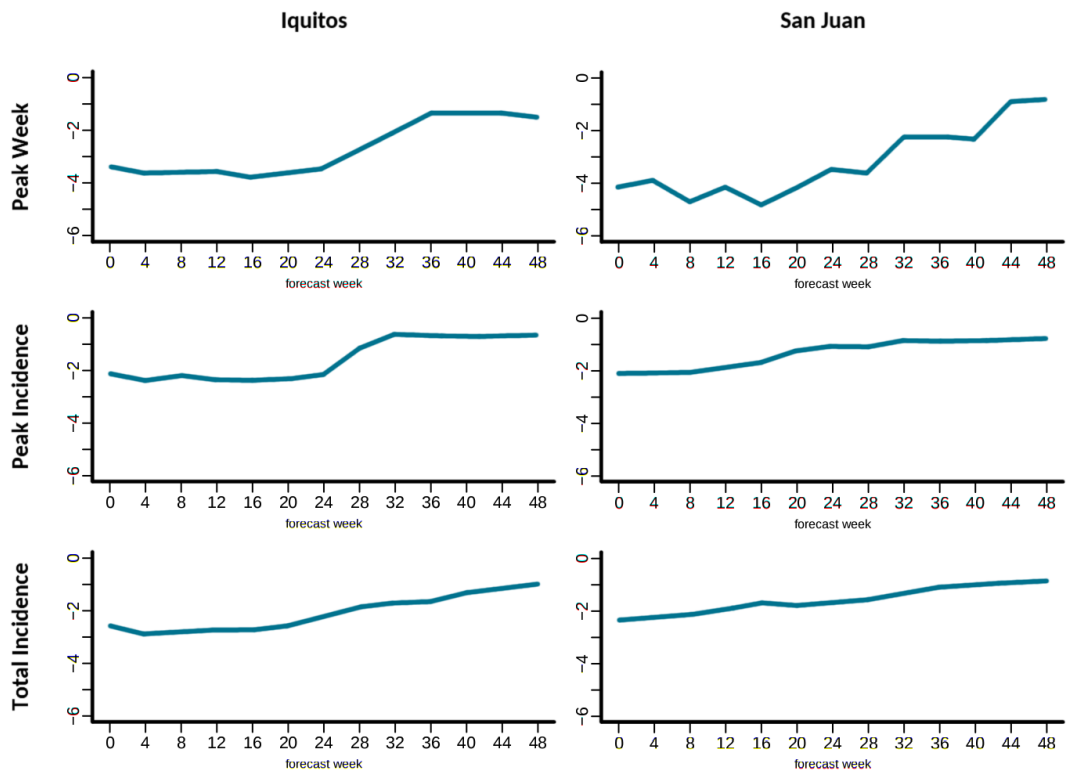


Figure 5.18: **Results of the dengue challenge in terms of log likelihood.** Weekly log likelihood, averaged over weeks of the 2009–2013 forecasting period, is shown for our team. Results are shown separately for each forecasting target and location.

### 5.6.2 DARPA Chikungunya Challenge

> Our forecasting system for the chikungunya challenge consisted solely of Epicast, which was implemented and maintained by David Farrow (myself). Roni Rosenfeld and Don Burke were both instrumental in soliciting the help of the participants with expertise in relevant fields.

The aim of the chikungunya challenge was unlike that of the influenza and dengue challenges in a number of ways. One way in which they differed has to do with the type of event being predicted. Whereas influenza and dengue have regular—and therefore somewhat predictable—outbreaks each year in their respective locations, chikungunya in 2014 had just been introduced to the Americas. As a result, this event was an *invasion*, not an epidemic in the more familiar sense of annual flu epidemics. For this reason, there is no historical case data available in the locations of interest, and the most similar datasets are for the invasion of other diseases in other locations. This alone is a significant obstacle for the application of statistical, data-driven forecasting methods. The other big difference in the chikungunya challenge is that no forecasts were required—only point predictions.

For the above reasons, we decided to apply our Epicast methodology to predict the chikungunya invasion, relying on expert opinion to build our aggregate prediction. We identified and invited around twenty experts in vector-borne viruses to participate in our Epicast methodology. Of these, twelve accepted the invitation and participated on at least one occasion during the challenge.

Once each month, from August, 2014 through January, 2015, I asked the expert participants to predict the cumulative weekly chikungunya case count in each of the 55 Pan American Health Organization (PAHO) locations through the end of February 2015; and throughout the challenge I gathered a total of 2,530 trajectories Figure 5.19.

A serious impediment to producing accurate predictions in the long term is the fact that errors in cumulative chikungunya forecasts accumulated over weeks, whereas errors in (non-cumulative) influenza and dengue forecasts were separated out across weeks. While it would have been trivial to convert a cumulative trajectory into a non-cumulative trajectory, the published counts which were defined to be ground truth are only available sporadically over time, preventing me from converting the true cumulative trajectory into a non-cumulative trajectory.

The increased difficulty of the task is reflected by a reduction in accuracy. At best (one week ahead), less than one in three predictions were within 10% of the actual value; and at worst (ten weeks ahead), over half of the predictions were off target by more than 50%. Even in such conditions, when comparing pair-wise absolute error between Epicast and each user, Epicast more frequently predicts closer to the true value than any individual user.

Epicast was not selected as one of the six chikungunya challenge winners [171], however we are told that it ranked in the top quartile (Q1) of submissions. In a more detailed breakdown of Epicast performance provided through communication with DARPA, we learned that the Epicast methodology was ranked Q1 in methodology, Q1 in predicting peak incidence, and Q1 on predicting incidence at the fifth monthly submission. On the remaining months, Epicast ranked Q3.

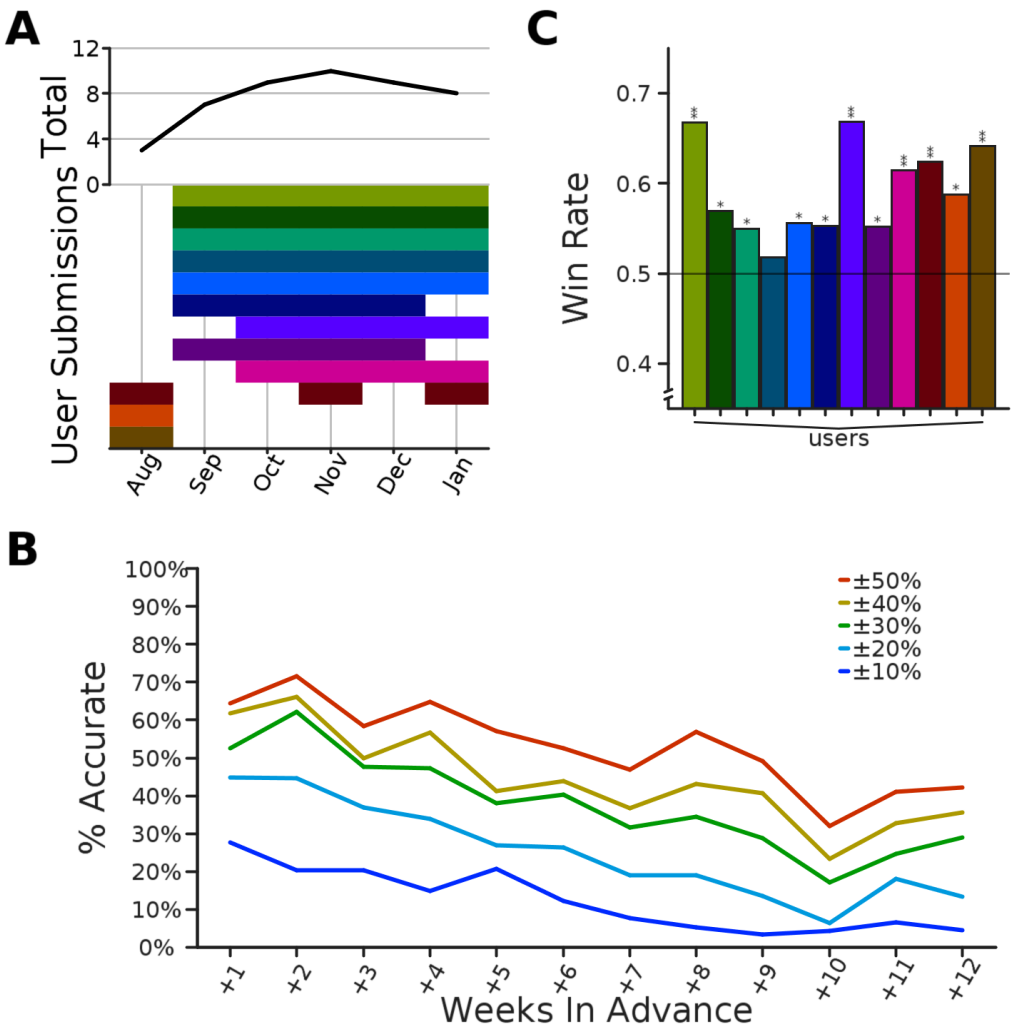


Figure 5.19: **Overview of Epicast chikungunya forecasts. (A)** Similar to Figure 5.6, participation is shown per month (top) and per expert (bottom). **(B)** As in Figure 5.7, percent of predictions within some range of the target value as a function of the number of weeks in advance that the prediction was made ($45 \leq n \leq 84$). **(C)** As in Figure 5.10, the fraction of instances where Epicast had lower absolute error than each individual participant, across all countries and weeks ($336 \leq n \leq 795$; Sign test; *: $p < 10^{-2}$; **: $p < 10^{-5}$).

## 5.7 Final considerations

### 5.7.1 Pros, Cons, and Caveats of human judgment in forecasting

For years, both humans and machines have been employed to tackle difficult prediction problems. The biases involved and the relative advantage of data-driven approaches are at least well documented [172, 173], if not well understood. I do not make the claim that human judgment is intrinsically more valuable or more capable than machines when making epidemiological forecasts, but I do posit that there is value in understanding the strengths in each approach and suspect that both can be combined to create a forecasting method superior to either approach alone. A similar discussion has recently taken place in the related field of numerical weather prediction [174].

There are several important limitations of the human judgment approach relative to purely data-driven methods that should be made clear. First, these results are only representative of a single flu season and a single chikungunya outbreak. This highlights one of the biggest shortcomings of this approach—collecting predictions is a tedious and time-consuming process. Unlike statistical methods which can be applied retrospectively to any outbreak, the approach here requires a significant amount of work from a large number of participants. Because of this I am unable to perform cross validation across seasons. Second, these results do not necessarily provide us with an improved understanding of epidemiological dynamics. In contrast, statistical methods can aim to learn from past data in order to better describe and model the epidemic process.

On the other hand, the human judgment approach does have unique advantages over purely data-driven systems. Humans have the innate and powerful ability to assimilate, with little to no effort, diverse data sources. An example of this is using news headlines, which we display within the Epicast interface, to inform predictions. Another advantage of human judgment is the ability to make reasonable predictions for events with little historical precedent, like the outbreak of a new disease or a disease invasion in a new location.

The task of predicting trajectories is not necessarily trivial, and I asked each of the participants to provide many such trajectories over quite a long period of time. There has been much work done to understand the ways in which crowd work can be optimized in terms of maximum benefit to participants [175], and I made every effort to achieve this goal. To minimize the overall amount of effort required and to streamline the process as much as possible I: allowed users to use their previously entered forecasts as a starting point; accepted any number of regional flu predictions (not requiring all eleven to be completed); reduced the entire process to one drag and one click per region, and sent URLs tailored with a unique identifier via email each week to bypass having to login. Additionally I tried to increase interest and participation by including a leader board of both weekly and overall high scores. I also had the competing objective of collecting the most informed forecasts from our users. To this end, I included a section of links to educational resources, and I embedded within each user's home page a Google news feed on the topic of "flu".

It was our hope that the number of participants would grow organically, for example through word of mouth and social media. Instead, we found it difficult to recruit new participants and to maintain participation throughout the flu season. The failure to achieve a true "crowd" is most

likely due to the tedium of the task, and I have considered ways to both reduce this tedium and to make the task more gratifying for participants. While I strove to design the user interface in a way that minimizes the level of effort required to input predictions, there is always room for further improvement. One option I considered, but did not implement because of the small number of participants, is to reduce workload by asking participants to provide a prediction only for a randomized subset of regions. Another option I considered, but did not completely implement due to time constraints, was gamification. This was partially implemented in the form of leader boards, but it would be difficult to provide a more immediate reward because of the inherent delay between prediction and revelation of true outcomes.

### 5.7.2 The state of the art

Epicast is first and foremost a bona fide approach to epidemiological forecasting; however, it also provides a valuable and intuitive baseline for assessing the state of the art. The focus of the various epidemiological forecasting challenges has been to build computational, data-driven forecasting methods, and the standards by which these systems have been judged are often arbitrary. One example of this is the performance of the uniform system as used in Figure 5.11; another example is the uniform prior used as a component in the dengue forecasting methodology. We hope that our data-driven methods outperform these baselines, but aside from these and other simple sanity checks, we are left without a sense of the overall value of our data-driven forecasts.

In this context, Epicast can be seen as another, and much more interesting, baseline: the best forecasts we would have at our disposal *without* the aid of computational methods. In this light, we see what we have to gain by developing computational approaches to forecasting, and we learn a more informative measure of the relative utility of such systems.

With this in mind, it is disappointing on some level that Epicast was the most accurate method in the 2014–2015 flu contest short-term targets. On the other hand, statistical systems were not so far off—some were slightly ahead in seasonal targets. I suspect that in short time some computational systems will overtake Epicast entirely, and when this happens it will represent significant progress in our understanding of, and in our ability to prepare for, disease outbreaks. For now, it appears that human judgment has a relative advantage over data-driven methods, at least in short-term forecasts.

### 5.7.3 A model of models

A natural evolution of systems such as those for epidemiological forecasting is the combination of human and statistical (machine) methods [175, 176]. The first question in such a project is whether human predictions should be given as input to statistical methods or whether the output of the statistical methods should be shown to humans for more informed predictions. In theory both directions are viable, and there are intuitive reasons for each. In support of the latter, people are naturally inclined to trust forecasts made by humans (or to distrust forecasts made by machines), a phenomenon known as algorithm aversion [173]. Supporting the former, on the other hand, is the observation that in many settings and in a variety of tasks, objective machine prediction is often superior to subjective human prediction [172, 177].

I have taken steps to explore both directions using the nowcasting system described in Chapter 4. In the current, ongoing iteration of Epicast (2015–2016 flu contest), I show a subset of participants the nowcast; and the nowcast system takes as input the 1 Week Lookahead from Epicast.

It has become apparent throughout the various forecasting challenges that mixtures of models generally produce better results than any of their component models. The methodology we applied in the dengue challenge is one such example, and the Empirical Bayes system for the 2015–2016 flu contest has evolved into a mixture of several methods. Combined approaches like these appear to be the way of the future.

# Chapter 6

# Conclusion

> There is a single light of science, and to brighten it anywhere is to brighten it everywhere.
>
> *Isaac Asimov*

## 6.1 Summary of contributions

Although modeling the epidemiological trends of influenza was relatively straightforward, modeling the evolutionary trends of influenza was a challenge for quite some time. One of the earliest models that was able to explain the more complete phylodynamic picture of influenza made use of an untested and theoretical immune component: generalized immunity. While it is not known with certainty that this mechanism exists in humans and is solely responsible for influenza's distinctive evolutionary dynamics, it at least appears plausible, based on a variety of empirical evidence. In Chapter 3 the main question I asked is this: if generalized immunity does in fact exist, how can it be most plausibly characterized? To answer that question, I mapped the parameter space of generalized immunity and compared simulated outcomes with those observed empirically for influenza A/H3N2. In addition to providing independent confirmation that a model with generalized immunity is capable of capturing influenza's full range of dynamics, I found that generalized immunity could plausibly be much weaker and longer lasting than previously hypothesized.

Beyond providing a more thorough characterization of generalized immunity, I also contribute in Chapter 3 a methodology for evaluating simulated outcomes with respect to some set of reference outcomes. Determining whether, and to what extent, a model-generated trajectory matches expected outcomes for influenza has been a challenge for as long as influenza has been modeled, and a large number of individual metrics have been conceived and applied to the task. Unfortunately though, no standard methodology has been suggested for such evaluation, and outcomes from each model have generally been studied qualitatively with disjoint sets of measurements and methods. More problematic though is a lack of statistical support for these methods, in part because the measures employed are typically considered only in isolation. I showed that many epidemiological and evolutionary measures can contribute simultaneously to

our confidence in the plausibility of generated trajectories in comparison to a set of reference values. I posit and test the hypothesis that a simulation is plausibly influenza-like, enabling the assignment of likelihoods to individual trajectories and of plausibility to specific parameterizations. Looking forward, it is my hope that similarly quantitative approaches will be used to lend statistical support to analyses of simulated dynamics of infectious diseases.

Having a reliable model of influenza's phylodynamics implies that we have, at least on some level, a working understanding of the processes that drive outbreaks. This is a critical requirement as we transition from explanations of the past to predictions of the future. Unfortunately one of the largest obstacles in making this transition is a lack of situational awareness due to both inherent and artificial shortcomings in traditional clinical surveillance. However, thanks to very recent and increasing usage of the internet as a tool for learning and communication, there is now available a tremendous amount of real-time digital surveillance data. In Chapter 4 I survey these data streams and then combine them to produce an estimate of influenza activity in the US in real-time. Data assimilation is nontrivial in this context; signals are noisy, heteroskedastic, variably correlated, intermittently available, and exist at varying temporal and geographic resolutions. Through application of a sensor fusion kernel derived from the Kalman filter, I use all available data to produce an optimal estimate of influenza activity within all US states—something that, to my knowledge, has never been previously attempted. By providing an accurate, timely, and high-resolution consolidated surveillance stream, this work directly facilitates epidemiological forecasting by addressing some of the most serious shortcomings of traditional surveillance.

Epidemiological forecasting is currently a very active area of research, and a large number of forecasting frameworks have been developed for a variety of infectious diseases, thanks in no small part to several contests and challenges sponsored by the US government. In Chapter 5 I focus on three such systems, two of which I contributed to significantly. The first of these, the Empirical Bayes system, takes a purely data-driven approach, producing probabilistic forecasts based on empirical trajectories of past epidemics. This system was applied with much success to forecasting both influenza and dengue, and its primary strength lies in a non-mechanistic approach to modeling and forecasting epidemic trajectories. The second system, Epicast, is based entirely on collective human judgment and represents a novel approach to epidemiological forecasting. In addition to being one of the winning systems in the 2014–2015 flu contest, Epicast was a strong competitor in forecasting chikungunya.

Many novel insights were gleaned through the development and application of Epicast and the other forecasting systems in Chapter 5. There is evidence that prediction accuracy is relatively consistent among human participants. Those who do well typically do well most rounds, whereas those who do poorly typically do poorly most rounds. It was unclear beforehand whether this would be the case—now we have strong evidence suggesting that it is. As previously discussed, this knowledge can be used, for example, to weight individuals or systems based on past performance to produce a better overall forecast. Until Epicast, the only way to evaluate data-driven forecasts was to compare those forecasts with other data-driven forecasts and with very simple and predefined baselines. Now we have something very valuable: a human baseline. To say that a computational forecast outperforms a computational baseline is not always very meaningful; but to show that a computational forecast outperforms the best forecasts that humans could otherwise produce is quite meaningful. In this sense Epicast provides an intuitive measure of the state of the art, and it appears that we are at a point in time when computational forecasting is

just beginning to surpass that of human judgment.

## 6.2 Future directions

Although there is some biological evidence in support of a generalized immune response against influenza, it has not yet been conclusively shown to exist in humans. A prominent alternative hypothesis instead suggests that ephocal evolution, in the absence of generalized immunity, is sufficient to constrain diversity. This epochal evolution was originally modeled using a neutral network genotype-to-phenotype map [47] and was later generalized in a model considering only the tempo of antigenic change [131]. Other hypotheses of influenza evolution posit an extremely limited set of antigenic phenotypes [120, 178] or suggest that antigenic evolution is canalized by human immunity [25]. As new explanations for influenza's distinctive dynamics arise, an important task will be to quantitatively and objectively validate and contrast results from each of these models. My simulator of influenza transmission and evolution [111, 112] could be used as a starting point for such comparisons.

Ultimately, however, these are questions concerning *biological* processes; and all of the answers so far have been *computational* in nature. Biological experiments must be performed at some point to provide a more consistent and empirically supported explanation for the role of human immunity in shaping influenza's phylodynamics. In Chapter 3 I describe much more precisely the biologically plausible ranges of strength and duration of generalized immunity in the hope that future work toward its empirical validation can take advantage of these estimates to significantly cull the experimental search space.

A sensor fusion approach to nowcasting in Chapter 4 has enabled the assimilation of a diverse set of digital surveillance sources that were previously considered either entirely in isolation [69, 79, 82] or were used at the very coarse intersection of space and time for which they were simultaneously available [85]. While this combined approach represents a significant improvement in our ability to estimate real-time disease incidence, its applications are intrinsically limited by the availability of digital surveillance data. Disease nowcasting continues to lag outside of the US, especially in areas with limited internet access. Even in well-developed areas with widespread internet access, it seems that using the English language is a requirement for many signals, as these systems have only been trained on English datasets so far. While some progress has been made towards expanding the geographic coverage of digital surveillance signals [81, 144], more work is needed to bring disease nowcasting in other locations up to par with that in the US. Similarly, disease nowcasting is most advanced for influenza, and more work is needed to bring nowcasting to other diseases.

Another interesting dimension to the nowcasting problem has yet to be explored. I attempt to estimate the current gold standard of flu activity, wILI; but wILI is itself a convolution of noisy constituent signals. wILI is inherently based on the clinical manifestation and human recognition of symptoms—it is syndromic surveillance. Unfortunately, a large number of pathogens can cause flu-like symptoms. Two very prevalent examples of this are rhinoviruses and the human respiratory syncytial virus. Additionally, flu is caused not by any single virus, but by different types (including A and B) and subtypes (including A/H3N2 and A/H1N1). These variants are indistinguishable by the broad and subjective definition of ILI. There are, however, sources of viral

identifying information; one example is the National Respiratory and Enteric Virus Surveillance System (NREVSS), which reports antigenic classification and genetic makeup of viral samples submitted within the US. With appropriate data (including no doubt that from NREVSS), it should be possible within my sensor fusion framework to nowcast not only wILI, but also the pathogen makeup of wILI. Appropriate data, however, is critical, and for syndromic data sources this will require capturing the subtle differences between diseases in terms of symptoms, host behaviors, demographics, timing, and geographic distributions.

Because epidemiological forecasting is still a relatively new endeavor, there are numerous directions for future work. Since their respective inception for the 2014–2015 and 2013–2014 flu contests, the Pinned Spline and Empirical Bayes systems of Chapter 5 have evolved significantly. One potential future direction is to combine both of these systems to provide a single data-driven forecasting framework. This would be especially helpful considering the relative benefits of each system; Pinned Spline made strong point predictions, especially on short-term targets, and Empirical Bayes excelled at producing reasonable distributional forecasts. A natural goal then is to develop a single combined system with all of the benefits, and none of the drawbacks, of both approaches. This is, in fact, one of the goals of the recently developed "Delphi-Stat" system.

There are many ways in which the Epicast method could be improved and extended. There is an important relationship between a prediction, and the level of confidence in that prediction, that I was unable to capture. I asked participants to give their best point predictions, but there was no way for them to communicate their level of confidence in those predictions. I made the implicit assumption that disagreement among user predictions implies lack of confidence, which is probably true to some extent; the inverse however—that uniformity in predictions implies high confidence—is clearly untrue. Consider as an example the case where everyone believes that next week's wILI has a 60% chance of staying the same as this week's wILI, resulting in all point predictions strongly concentrated on the same wILI, and the distributional spread being very narrow, in contrast with the participants' beliefs. It would be ideal to collect from each user a more informative measure of their confidence. Second, an adaptive weighting scheme could be used to improve forecasting accuracy, similar in spirit to the way user recommendations and rankings are weighted, increasing accuracy in those settings [179, 180]. Preliminary results discussed in Chapter 5 suggest that some participants may consistently be more or less accurate than other participants, and an adaptive weighting scheme may benefit the overall forecast. Finally, there is the issue of sample size in terms of the number of participants in the Epicast "crowd". Forecasts from our relatively small crowd produced remarkably accurate forecasts; it is desirable, however, to have as large a crowd as possible. One way to achieve this may be, for example, through gamification.

## 6.3 Final thoughts

The subject of this thesis is the past, present, and future of influenza, and to a lesser extent, that of infectious diseases in general. But there has been a recurring theme throughout: combining all available evidence to accomplish a given task. This manifested in Chapter 3 by way of assessing plausibility of outcomes with respect to a set of empirically-derived targets. In my approach, I used Mahalanobis distance to combine evidence across epidemiological and evolutionary targets

which would have otherwise provided eight separate, likely correlated, and potentially conflicting, figures of merit for each parameterization of generalized immunity. Combining evidence is *the* major goal of Chapter 4. The primary challenge there, for which I used the sensor fusion kernel of the Kalman filter, was to assimilate signals with varying noise, resolution, and availability. This theme appeared in disguise throughout Chapter 5. In Pinned Spline and Empirical Bayes, we combine an enormous number of individual trajectories to build a single forecast; in comparing Epicast with the other methods, I combine accuracy across all weeks, targets, and regions to determine overall accuracy; and in Epicast, I combine point predictions from individual users to produce a full probabilistic forecast—all of these examples in some way make use of the multivariate normal likelihood. Data assimilation problems appear in a variety of settings, and a working knowledge of the tools and methods for handling these situations can be very powerful.

The methods I developed in this thesis are specifically applied to influenza, dengue, and chikungunya, but it is interesting to consider how they can be generalized for use in other settings. The method for determining plausibility in Chapter 3 could be used, for example, to determine the plausibility of other explanations of influenza's phylodynamics—or the plausibility of any phenomenon that can be simulated but not directly measured, and for which indirect empirical evidence is available. I used digital surveillance assimilation in Chapter 4 to estimate influenza activity, but I can imagine using other digital data streams to estimate changing political sentiment in real-time. I explored two particular generalizations in Chapter 5 by showing how our three forecasting frameworks could be applied to forecast dengue and chikungunya in tropical America. It is also worth noting that the Empirical Bayes system, at least, operates entirely on *trajectories*; given historical wILI trajectories and the current, partial wILI trajectory, it will forecast a new wILI trajectory—it is easy to imagine what it might do with, say, hourly temperature readings or daily exchange rates in place of wILI. In this sense, Empirical Bayes (and to some extent, Pinned Spline) can be seen as a general strategy for time series forecasting of recurring events.

For all of our advances and achievements as a society in past centuries, influenza and other infectious diseases continue to burden—and even threaten, in the case of pandemic. It is my lofty hope that one day influenza will no longer be as ubiquitous within human populations as it is today. This thesis is, I hope, one infinitesimally small step in that direction.

# Appendix A

# Epidemiological forecasting: a spline regression approach

The initial spline idea was due to Robert Tibshirani; Ryan Tibshirani, Roni Rosenfeld, and Sangwon Hyun also contributed significantly to the implementation of this idea. The spline framework was one of our entries in CDC's 2014–2015 flu forecasting contest—this is the version described below. In parallel, Sangwon Hyun and others used this framework to assess the risk of a dengue outbreak during the 2014 world cup in Brazil [181]. Later, Ryan Tibshirani and David Farrow (myself) implemented a modified version of the spline framework as a component of a combined entry in the OSTP dengue challenge. A manuscript for a forthcoming publication of the Pinned Spline system is currently being prepared by Xiaotong Suo and others.

## A.1 Intuition

The Pinned Spline approach was originally motivated by Epicast (Chapter 5). Consider the forecasting interface as shown for example in Figure 5.4. Participants are shown the most up to date—but incomplete—wILI trajectory for the current season, and they are also shown the wILI trajectories of past seasons. One prediction strategy is to draw a smooth continuation of the observed trajectory into the future such that the predicted trajectory matches the mean of past trajectories. This strategy can be summarized with the following goals:

- Start with the partially observed trajectory of the current season. wILI on past weeks is already known, up to reporting lag and revisions due to backfill; just report what has been observed.
- Predict wILI on future weeks as wILI on the same weeks averaged across past seasons. Use what is known about the shape of flu epidemics to predict wILI on the remaining weeks of the current season.
- Smoothly interpolate between predictions of the past and of the future.

These goals can be accomplished with the following two steps. First, use regression with a set of cubic basis *splines* to find a smooth trajectory that spans the entire season, matching

known wILI in the past and expected wILI in the future. Second, *pin* down the portion of the smooth trajectory corresponding to past weeks, replacing smoothed values with actual, known wILI values. This is the essence of the Pinned Spline forecasting method.

## A.2 Regression Splines

Splines and basis regression are advanced topics outside of the scope of this thesis. However, I briefly summarize the main ideas below. I direct the interested reader to a primer on spline regression [182] and to these textbooks for a much more thorough treatment of the subject [183, 184].

Splines are piecewise polynomial functions, named after the tool used by shipbuilders to produce wooden boards with smooth curves spanning fixed points. Basis splines, or "B-splines", are a class of splines that can be used in a regression setting to fit a smooth (maximally differentiable) curve through a set of observed points. A B-spline is defined by two parameters: the order, $m$, and the number of interior knots, $N$. The degree, $n$, of each of the spline's constituent polynomials is $m - 1$, and the total number of fixed points, including two endpoints, is, $N + 2$. To find the B-spline that best fits the data, a set of spline basis functions are used as regression covariates; the resulting B-spline is simply a linear combination of the individual basis splines. The spline basis functions for a cubic (order four, degree three) B-spline with ten interior knots, and also the best B-spline fit of this basis to sample wILI, are shown in Figure A.1.

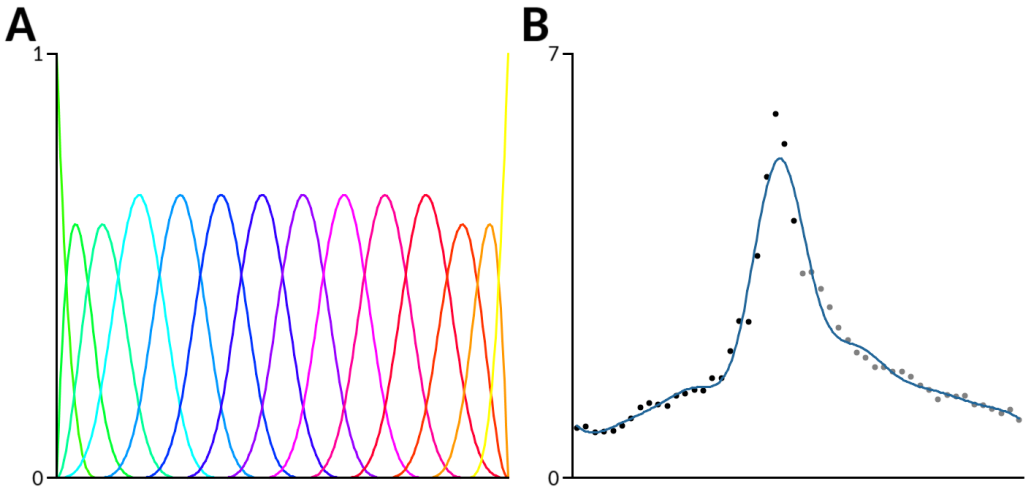


Figure A.1: **Spline basis functions and best fit B-spline. (A)** Basis functions for B-splines ($m = 4$, $N = 10$). **(B)** Observed wILI for 25 weeks of the 2014–2015 season (black points); target wILI for the following 25 weeks (grey points); regression fit of spline basis in panel A to wILI—a B-spline (blue line).

Given the spline basis, partially observed wILI, and mean wILI on future weeks, ordinary

least squares (OLS) regression is used to find the best fit B-spline to the data. The observed portion is then replaced with known wILI (pinning), and predictions are made based on measuring the desired targets (e.g. Peak Week, Peak Height) on the resulting curve.

## A.3 From point prediction to distributional forecast

The spline method described above produces a single trajectory—a prediction. As discussed in Chapter 5, it is much more desirable (and was required for the flu contest) to produce a distributional forecast. This is done in the Pinned Spline method by bootstrapping [184].

In the case of flu, the signal we want to forecast, wILI, is inherently noisy. This can be exploited to build a large set of predicted trajectories from which a forecast can then be derived. The general method is as follows. Smooth the wILI trajectory of all of the past seasons, for example using trendfiltering [170] as is done with the Empirical Bayes system (Figure 5.2). The result of this smoothing is two pieces of information for each curve: a smooth curve that interpolates the original wILI values and an estimate of the noise level, $\tau$. Next, randomly generate pseudo-observations by taking each smoothed past trajectory and adding Gaussian noise with zero mean and scale equal to $\tau$. With these new trajectories, run the original Pinned Spline procedure.

When this process has been repeated many times, the result is a set of B-splines, each fit to a random, but plausible, wILI trajectory. Similar to Epicast and Empirical Bayes, the point prediction of each target is defined to be the median value of the target measured on all splines, and the distributional forecast is the distribution of target values measures on all generated splines.

While this method is conceptually straightforward and simple to implement, the forecasts it generated during the 2014–2015 flu season were not very accurate. In particular, this method appears to suffer from general overconfidence. This was originally mitigated by blending the posterior distribution with a uniform distribution over all outcomes, but this is not a permanent solution. After the contest ended, one of our main goals was to improve the methodology to provide better, less overconfident, distributional forecasts.

It should be noted, however, that the Pinned Spline point predictions were quite accurate. In fact, accuracy of the Pinned Spline system approached that of Epicast in predicting the four Lookahead targets, and Pinned Spline was generally more accurate on short-term point predictions than our Empirical Bayes system.

# Appendix B

# The Archefilter: combined nowcasting and forecasting

## B.1 Overview

The Archefilter is a framework that I designed to predict wILI on any week, given tentative wILI values of past weeks and wILI of past seasons. It draws inspiration from our (significantly more complex) statistical forecasting systems, Empirical Bayes (Subsection 5.3.1) and Pinned Spline (Appendix A). The Archefilter produces forecasts on a weekly basis for the US nationally and for all HHS and census regions.

## B.2 The Archetype

A variety of approaches have been taken for forecasting influenza, including use of compartmental models, agent-based simulations, statistical methods, and a system based on collective human judgment. Compartmental models have the advantage of mathematical simplicity, but are generally unable to recapitulate the empirical shape of influenza trajectories; agent-based simulations better capture the epidemic trajectory, but make many strong mechanistic assumptions; statistical methods make little to no mechanistic assumptions, but require a large amount of data to estimate parameters; and human judgment methods produce very accurate short-term forecasts, but require a large investment of human time and effort. I attempt to capture the benefits, and avoid the drawbacks, of these approaches within the ARCH framework—with a focus on nowcasting and short-term forecasting. Based on the assumption that future epidemics will look something like past epidemics, the ARCH kernel consists of a description of the canonical shape of an influenza epidemic—an archetype. The resulting archetype is: simple, parameterized only by timing and magnitude; lightweight, requiring no simulation; empirical, making no mechanistic assumptions; and accurate, especially in the short-term. This methodology, based only on empirical data, is in theory directly applicable to other diseases with regularly occurring outbreaks, even if there is only a small amount of available historical data.

The archetype influenza trajectory is built through the following process, illustrated in Figure B.1:

1. Split the historical wILI trajectory into individual epidemic trajectories spanning from week 30 of one first year to week 29 of the next year.
2. Model and remove excess wILI on holiday weeks.
3. Smooth trajectories (for example, moving average, kernel smoother, or trend filtering). For simplicity, I use a Gaussian kernel smoother with a bandwidth of 2 weeks.
4. Rotate both the original and the smoothed trajectories such that the peak falls in the center of the season (week 3).
5. Calculate the week-wise mean wILI for both the original and smoothed trajectories.
6. Interpolate between original mean (middle of the season) and the smoothed mean (early and late in the season); I use the Hann function for this purpose.

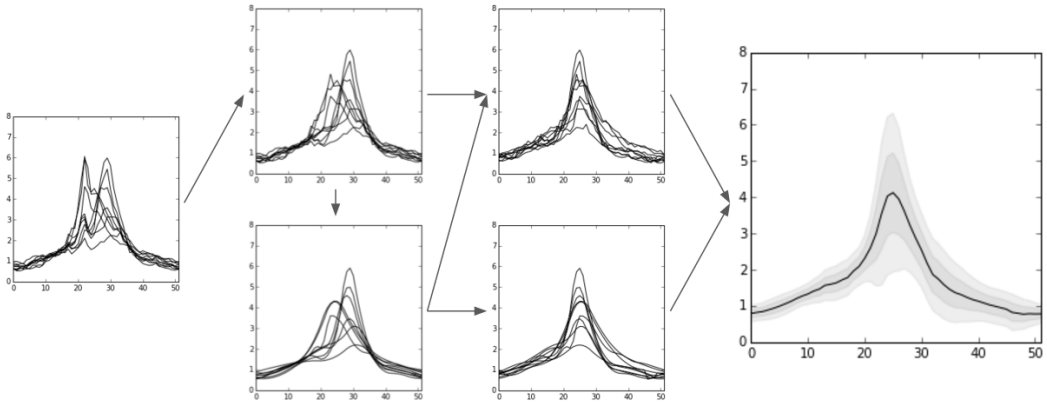

Figure B.1: **Graphical representation of the construction of the US national influenza archetype.** Far left: wILI trajectory for past seasons. Middle left: original (top) and smoothed (bottom) trajectories after attempting to remove the holiday effect. Middle right: original (top) and smoothed (bottom) trajectories aligned such that the peak falls in the middle of the season. Far right: the archetype (black) and its credible interval (gray).

As discussed in Chapter 4, the holiday effect is a term we use to describe the aberrant increase in reported wILI during the major holiday season. Unlike the additive approach in SAR3, here I find a multiplicative constant for the last 3 weeks of the old year and the first week of the new year. To determine this constant, I use derivative-free optimization [158] to find the wILI multipliers that achieve an approximately constant first derivative of wILI on these weeks. In other words, I find values with which to multiply wILI such that wILI on the holiday weeks is as near as possible (in a least-squares sense) to a linear interpolation of wILI between the week preceding the holiday to the week proceeding the holiday. The resulting trajectory—the archetype—is a process model for wILI, similar in some respect to an SIR trajectory. The difference, however, is that the archetype is an *empirical* description the the flu process rather than a mechanistic one. An archetype curve is constructed separately for the 19 US regions and nationally (20 total).

## B.3 The Archefilter

Defining the archetype trajectory is just the first half of the ARCH method; the second half is to produce a prediction of future wILI. To do this, I find a *transformed* version of the archetype curve that best explains the partially observed wILI trajectory of the current season. These transformations are a shift in time (rotation, or circular shift, along the x-axis) and a scale in magnitude (a multiplication on the y-axis). To assess how well a particular transformed instance of the archetype explains observed wILI, I use the multivariate normal likelihood as explained below.

The best-fit instance of the archetype will fit very closely to observed wILI—especially wILI on recent weeks. It will also fit reasonably well to the mean of past seasons; this is to prevent selection of an archetype instance that over-fits observed wILI at the cost of making unrealistic predictions about future wILI. An example of such fits are shown in Figure B.2. The goal of forecasting, however, is not to produce a singe possible future, but to produce a distribution over possible futures. This is done in the ARCH system the same as in the Empirical Bayes system: by building a posterior distribution of curves and reporting the statistics of that distribution.

First, a prior distribution must be created. I define this beforehand as the space of all archetype curves that have been transformed by shifts of up to 10 weeks in either direction (uniformly) and scales from 33% to 300% (uniformly). To get from the prior to the posterior, I need to more precisely define the likelihood function. As in Empirical Bayes, I use the multivariate normal likelihood. On runtime week $wk$, with observed (past) and imputed (future) wILI values ($y \in \mathbb{R}^{52}$), the weight assigned to any sample from the prior ($s \in \mathbb{R}^{52}$) is:

$$\text{weight}^{-1} \propto (y-s)^T \Sigma^{-1} (y-s).$$

Here, however, $\Sigma$ is carefully constructed to meet certain expectations. These are: (a) the recent past (5 weeks) should match observed wILI very closely, (b) the more distant past should match observed wILI, up to the uncertainty of backfill, and (c) the future should match the week-wise mean wILI of past seasons, up to variability in historical wILI. I define each of these more concretely next and show how these ideas are encoded in $\Sigma$. To begin with, $\Sigma = I$ $(52 \times 52)$.

To meet condition "a", I multiply rows of $\Sigma$ by $10$ everywhere *except* for the rows corresponding to the most recent 5 weeks. That is, rows $1..(wk-5)$ and rows $wk..52$. This reflects my strong desire to match the recent past, relative to the distant past and the future. I do this because I want to focus primarily on nowcasting and short-term forecasting.

To meet condition "b", I divide rows of $\Sigma$ by the empirical variance of backfill in the region. I compute the backfill variance (separately for each region) by comparing finalized wILI values to preliminary values. I do this separately for each "lag" time: the age, in weeks, of preliminary wILI. I do this because backfill variance decreases as lag time increases. In the interest of running time, I limit this to lags of 1–10 weeks; anything older than 10 weeks I assume has the backfill variance of 10 weeks. Finally, I divide the rows of $\Sigma$ which correspond to the past (that is, rows $1..(wk-1)$) by the respective variance of backfill. My reasoning here is that observed values in the recent past—which are subject to the largest backfill—should not be relied on as heavily as observed values in the more distant past. In other words, avoid overly penalizing archetypes that deviate from observed values *if* that deviation can be explained by backfill.

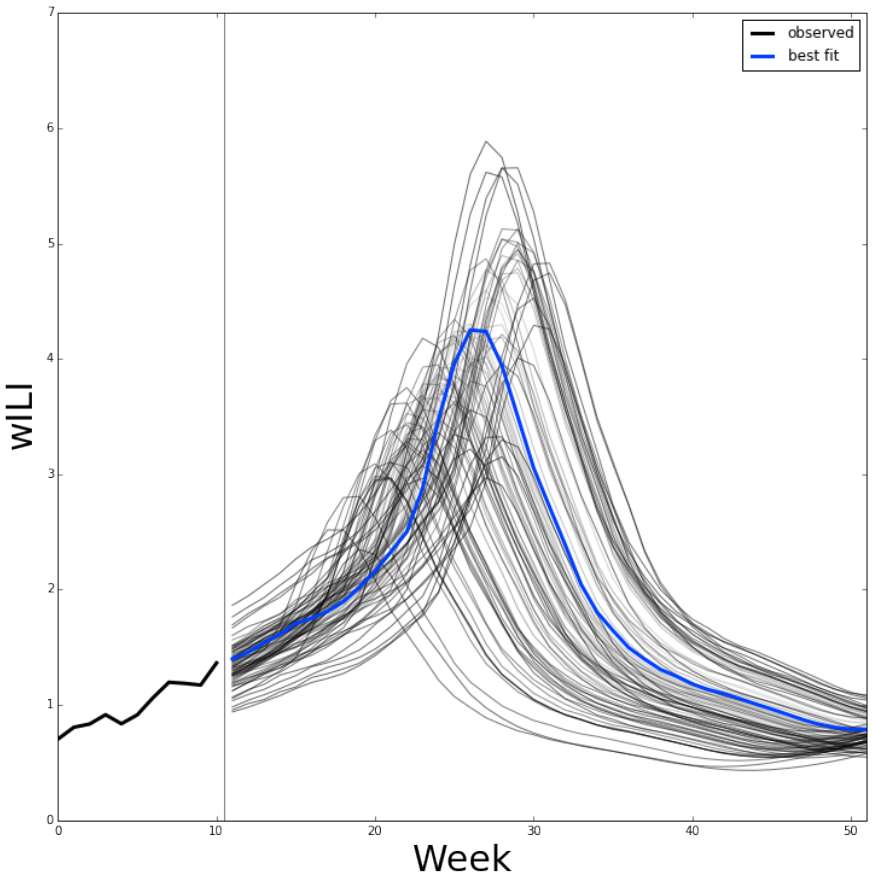


Figure B.2: **Archetpye transformations and best fit.** Partially observed wILI for the current season is shown in black. Archetype curves, transformed by randomly selected shifts in time and scales in height, are shown as thin grey curves; better fits are darker, and worse fits are lighter. The single best fit instance of a transformed archetype is shown in blue.

To meet condition "c", I divide rows of $\Sigma$ by the empirical week-wise variance of wILI of past seasons in the region. For all future weeks ($i \in [wk..52]$), I calculate the variance of wILI curves ($w$) from each of the $n(=11)$ past non-pandemic seasons ($s \in [2003..2008, 2010..2014]$), as (using array index notation):

$$\text{var}_i = \frac{1}{n}\sum_{j=1}^{n}(w_{s_j}[i] - \bar{w[i]})^2.$$

Then, as before, I divide the rows of $\Sigma$ which correspond to the future (that is, rows $wk..52$) by the respective weekly variance of historical wILI, $\text{var}_i$. My reasoning is that some parts of the season are inherently more variable than others. Namely, the middle of the season is highly variable, but the end of the season is relatively static. I want to penalize deviations from the norm, under the assumption that the current season will turn out something like past seasons.

For example, it is difficult to say that 3 wILI units above the mean is unrealistic in January; however, 3 wILI units above the mean in April is quite unrealistic based on the non-pandemic seasons we have observed so far. In other words, I want to constrain my posterior samples by what could plausibly happen in the future, and one way to describe what is plausible is by mean and variance of wILI on past seasons.

Having defined the prior, $\Sigma$, and the likelihood, I now describe the process of sampling from the posterior. First, I scan a grid over shift and scale parameter space. I then normalize the resulting weight values to form a proper probability mass function (PMF) over parameter space, as illustrated in Figure B.3.

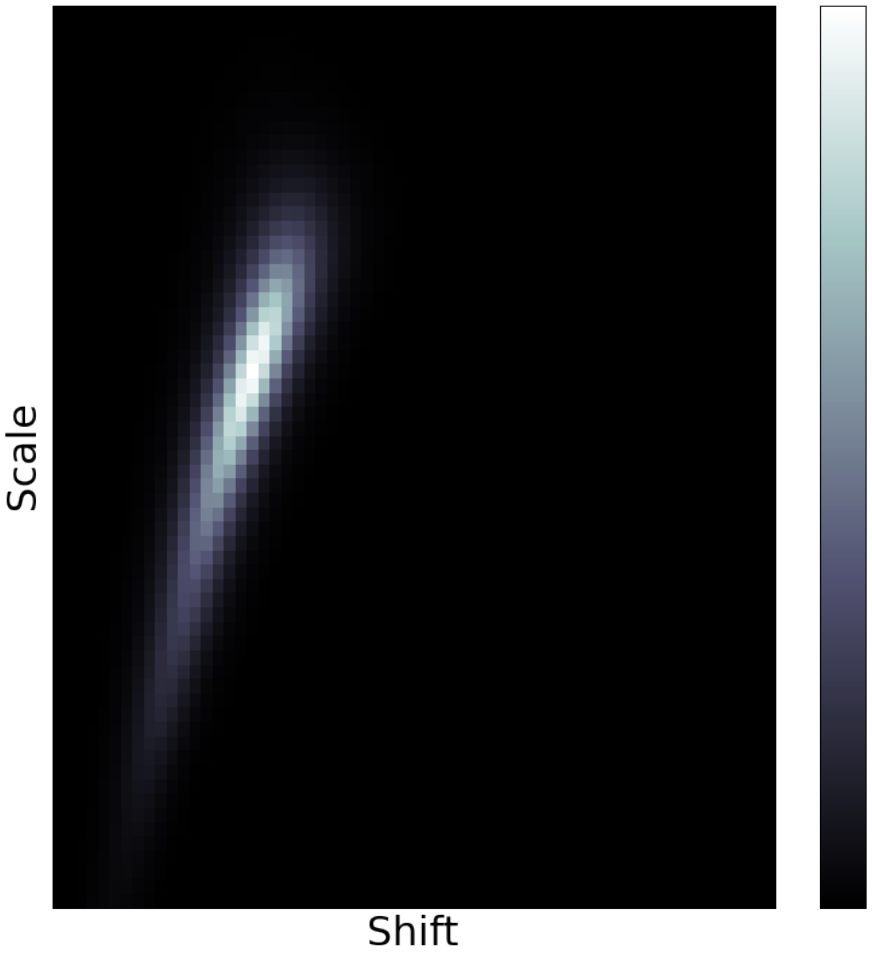


Figure B.3: **Parameter space of archetypes fit to wILI.** Archetypes are instantiated with shifts ranging from -10 to +10 weeks and scales ranging from 33% to 300%. Each instance is compared to observed wILI (in the past) and average wILI of past seasons (in the future), with weight determined by multivariate normal likelihood. Weights are normalized to form a proper PMF, and this is plotted on the grid shown. Weights are colored such that good fits are lighter and poor fits are darker.

Next, I draw samples from this PMF; each sample describes a particular instantiation of the archetype. I store each sample and its associated weight. Finally, I report distributions over target values. For the CDC flu contest, these were: Peak Week, Peak Height, and 1–4 Week Lookaheads (see Chapter 5 for much more information). For point predictions, I report the weighted median of the target values measured on each posterior sample. For distributional forecasts, I report a Gaussian distribution with mean and variance calculated on weighted target values of each sample. To prevent overconfidence and to allow for surprises, I use two post-processing methods to spread the probability among the reported bins. First, I smooth the distributional probability bins with Gaussian kernel smoother (bandwidth = $1$ bin). Second, I blend the bins with a uniform distribution with weight chosen such that no bin is assigned probability less than $0.002$.

## B.4 Use within the Kalman filter

The archetype was originally intended to be a process model for use in the Kalman filter (KF). Usage of the KF was however entirely obviated and superseded by the development of my sensor fusion framework in Chapter 4. Though I no longer use this method, I describe it here for completeness.

The problem at hand is one of *tracking*: given a model of the flu process and noisy estimates of wILI, produce an optimal estimate of finalized wILI. With the archetype as a process model and TWTR, WIKI, and SAR3 (see Chapter 4) as noisy estimates of wILI, I have everything needed for using the KF.

There is one important caveat, however: the "process" (archetype) is *highly* nonlinear. One solution is to break away from the traditional formulation of the KF and use a relaxed version that can handle nonlinearities: the Unscented Kalman filter (UKF) [185]. The UKF provides a very general interface through which the process can be queried. To fit this interface, the process must take as input some set of parameters and produce as output a state estimate (wILI in this case). This is exactly what the archetype does; given shift, scale, and current week, the wILI estimate is the point on the shifted and scaled archetype instance one week in the future.

I use the UKF as implemented in the python software package filterpy [186, 187]. Usage is straightforward; on any given week, I provide these things:

1. Prior over states (mean $x$ and variance $P$)
2. Process variance ($Q$)
3. Measurement covariance ($R$)
4. Measurements (TWTR, WIKI, and SAR3; $z$)
5. Process model (archetype)

I define the prior mean $x$ to be the most recently published value of wILI. I define the prior variance $P$ to be the backfill variance, as previously discussed, at a lag of one week. I roughly estimate, based on intuition and visual inspection, the process variance $P$ to be 0.25 (standard deviation of 0.5 wILI). Similarly, I roughly estimate, based on intuition and visual inspection, the measurement covariance $Q$ to be the diagonal matrix of variances for TWTR, WIKI, and SAR3: 0.49, 0.25, and 0.25 (standard deviations of 0.7, 0.5, and 0.5 wILI), respectively. $z$ is

simply the vector of TWTR, WIKI, and SAR3 readouts for the current week. The process model is the previously described archetype. Note that the archetype is the epidemic curve *model* and the archefilter is the *forecasting system*.

With all pieces in place, I invoke the UKF to produce an approximately optimal estimate of the posterior state: wILI on the next week—a nowcast. With the nowcast in hand, I use the archefilter to build a forecast—a distribution over possible futures instead of a single possible future. Thanks to the nowcast, this forecast has the advantage of "seeing" one additional data point.

Going forward, the tracking (nowcasting) problem is essentially solved, at least for the purposes of the archefilter. Instead of using the UFK to manually produce a nowcast, I can simply use the nowcast that the sensor fusion framework produces. This decoupling of nowcasting and forecasting is beneficial for a number of reasons. For example, the two problems, though they may appear similar on the surface, are quite different; it makes sense then from an engineering standpoint to split the tasks into separate projects. Much more importantly, the nowcast not only is useful for the archefilter, but also is helpful for forecasting systems in general and is potentially suitable for distribution to the general public.

# Bibliography


[1] World Health Organization. World development report 1993: investing in health. *Commun Dis Rep CDR Wkly*. 1993;3:137.

[2] Armstrong GL. Trends in Infectious Disease Mortality in the United States During the 20th Century. *JAMA*. 1999 jan;281(1):61.

[3] Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJ. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *The Lancet*. 2006 may;367(9524):1747–1757.

[4] Pinner RW. Trends in Infectious Diseases Mortality in the United States. *JAMA*. 1996 jan;275(3):189.

[5] Cox NJ, Subbarao K. Global epidemiology of influenza: past and present. *Annual review of medicine*. 2000 Jan;51:407–21.

[6] Influenza (Seasonal). World Health Organization; 2016. Available from: `http://www.who.int/mediacentre/factsheets/fs211/en/`. Visited on: 2016-03-25.

[7] Impagliazzo A, Milder F, Kuipers H, Wagner MV, Zhu X, Hoffman RMB, van Meersbergen R, Huizingh J, Wanningen P, Verspuij J, de Man M, Ding Z, Apetri A, Kükrer B, Sneekes-Vriese E, Tomkiewicz D, Laursen NS, Lee PS, Zakrzewska A, Dekking L, Tolboom J, Tettero L, van Meerten S, Yu W, Koudstaal W, Goudsmit J, Ward AB, Meijberg W, Wilson IA, Radošević K. A stable trimeric influenza hemagglutinin stem as a broadly protective immunogen. *Science*. 2015;349(6254):1301–1306.

[8] Carrat F, Flahault A. Influenza vaccine: the challenge of antigenic drift. *Vaccine*. 2007;25(39):6852–6862.

[9] Horimoto T, Kawaoka Y. Pandemic threat posed by avian influenza A viruses. *Clinical microbiology reviews*. 2001;14(1):129–149.

[10] Russell CA, Fonville JM, Brown AEX, Burke DF, Smith DL, James SL, Herfst S, van Boheemen S, Linster M, Schrauwen EJ, Katzelnick L, Mosterín A, Kuiken T, Maher E, Neumann G, Osterhaus ADME, Kawaoka Y, Fouchier RAM, Smith DJ. The Potential for Respiratory Droplet–Transmissible A/H5N1 Influenza Virus to Evolve in a Mammalian Host. *Science*. 2012;336(6088):1541–1547.

[11] Lamb R, Krug R. Orthomyxoviridae: the viruses and their replication, p 1487–1531. vol. 1. LWW; 2001.

[12] Overview of Influenza Surveillance in the United States. Centers for Disease Control

and Prevention; 2015. Available from: http://www.cdc.gov/flu/weekly/overview.htm. Visited on: 2016-01-22.

[13] WHO surveillance case definitions for ILI and SARI. World Health Organization; 2014. Available from: http://www.who.int/influenza/surveillance_monitoring/ili_sari_surveillance_case_definition/en/. Visited on: 2016-02-25.

[14] Drake JW, Charlesworth B, Charlesworth D, Crow JF. Rates of spontaneous mutation. *Genetics*. 1998;148(4):1667–1686.

[15] How Flu Spreads. Centers for Disease Control and Prevention; 2013. Available from: http://www.cdc.gov/flu/about/disease/spread.htm. Visited on: 2016-02-25.

[16] Influenza National and Regional Level Graphs and Data. Centers for Disease Control and Prevention; 2015. Available from: http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html. Visited on: 2016-01-22.

[17] MMWR Week Fact Sheet. Centers for Disease Control and Prevention; 2015. Available from: http://wwwn.cdc.gov/nndss/document/MMWR_Week_overview.pdf. Visited on: 2016-01-20.

[18] Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford Ja, Holmes EC. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science (New York, NY)*. 2004 Jan;303(5656):327–32.

[19] Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus ADME, Fouchier RaM. Mapping the antigenic and genetic evolution of influenza virus. *Science (New York, NY)*. 2004 Jul;305(5682):371–6.

[20] Hay A, Gregory V, Douglas A, Lin Y. The evolution of human influenza viruses. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*. 2001 Dec;356(1416):1861–70.

[21] Wolf YI, Viboud C, Holmes EC, Koonin EV, Lipman DJ. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biology direct*. 2006 Jan;1:34.

[22] Bush RM, Fitch WM, Bender Ca, Cox NJ. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Molecular biology and evolution*. 1999 Nov;16(11):1457–65.

[23] Shih ACC, Hsiao TC, Ho MS, Li WH. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proceedings of the National Academy of Sciences of the United States of America*. 2007 Apr;104(15):6283–8.

[24] Nelson MI, Simonsen L, Viboud C, Miller Ma, Taylor J, George KS, Griesemer SB, Ghedin E, Ghedi E, Sengamalay Na, Spiro DJ, Volkov I, Grenfell BT, Lipman DJ, Taubenberger JK, Holmes EC. Stochastic processes are key determinants of short-term evolution in influenza a virus. *PLoS pathogens*. 2006 Dec;2(12):e125.

[25] Bedford T, Rambaut A, Pascual M. Canalization of the evolutionary trajectory of the

human influenza virus. *BMC biology*. 2012 Jan;10(1):38.

[26] Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, Russell CA, Smith DJ, Pybus OG, Brockmann D, Suchard MA. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathog*. 2014;10(2):e1003932.

[27] Koelle K, Ratmann O, Rasmussen Da, Pasour V, Mattingly J. A dimensionless number for understanding the evolutionary dynamics of antigenically variable RNA viruses. *Proceedings Biological sciences / The Royal Society*. 2011 Dec;278(1725):3723–30.

[28] Spicer CC, Lawrence CJ. Epidemic influenza in Greater London. *The Journal of hygiene*. 1984 Aug;93(1):105–112.

[29] Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, Gregory V, Gust ID, Hampson AW, Hay AJ, Hurt AC, de Jong JC, Kelso A, Klimov AI, Kageyama T, Komadina N, Lapedes AS, Lin YP, Mosterin A, Obuchi M, Odagiri T, Osterhaus ADME, Rimmelzwaan GF, Shaw MW, Skepner E, Stohr K, Tashiro M, Fouchier RaM, Smith DJ. The global circulation of seasonal influenza A (H3N2) viruses. *Science (New York, NY)*. 2008 Apr;320(5874):340–6.

[30] Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, Gregory V, Gust ID, Hampson AW, Hay AJ, Hurt AC, de Jong JC, Kelso A, Klimov AI, Kageyama T, Komadina N, Lapedes AS, Lin YP, Mosterin A, Obuchi M, Odagiri T, Osterhaus ADME, Rimmelzwaan GF, Shaw MW, Skepner E, Stohr K, Tashiro M, Fouchier RAM, Smith DJ. Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine*. 2008;26:D31–D34.

[31] Bedford T, Riley S, Barr IG, Broor S, Chadha M, Cox NJ, Daniels RS, Gunasekaran CP, Hurt AC, Kelso A, Klimov A, Lewis NS, Li X, McCauley JW, Odagiri T, Potdar V, Rambaut A, Shu Y, Skepner E, Smith DJ, Suchard MA, Tashiro M, Wang D, Xu X, Lemey P, Russell CA. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*. 2015;.

[32] Finkelman BS, Viboud C, Koelle K, Ferrari MJ, Bharti N, Grenfell BT. Global patterns in seasonal activity of influenza A/H3N2, A/H1N1, and B from 1997 to 2005: viral coexistence and latitudinal gradients. *PloS one*. 2007 Jan;2(12):e1296.

[33] Shaman J, Kohn M. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences of the United States of America*. 2009 Mar;106(9):3243–8.

[34] Shaman J, Pitzer VE, Viboud C, Grenfell BT, Lipsitch M. Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS biology*. 2010 Feb;8(2):e1000316.

[35] Grais R, Ellis JH, Kress A, Glass G. Modeling the spread of annual influenza epidemics in the US: The potential role of air travel. *Health care management science*. 2004;7(2):127–134.

[36] Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, Grenfell BT. Synchrony,

waves, and spatial hierarchies in the spread of influenza. *science*. 2006;312(5772):447–451.

[37] Brownstein JS, Wolfe CJ, Mandl KD. Empirical evidence for the effect of airline travel on inter-regional influenza spread in the United States. *PLoS Med*. 2006;3(10):e401.

[38] Chowell G, Miller Ma, Viboud C. Seasonal influenza in the United States, France, and Australia: transmission and prospects for control. *Epidemiology and infection*. 2008 Jun;136(6):852–64.

[39] Nukiwa-Souma N, Burmaa A, Kamigaki T, Od I, Bayasgalan N, Darmaa B, Suzuki A, Nymadawa P, Oshitani H. Influenza transmission in a community during a seasonal influenza A(H3N2) outbreak (2010-2011) in Mongolia: a community-based prospective cohort study. *PloS one*. 2012 Jan;7(3):e33046.

[40] Gog JR, Grenfell BT. Dynamics and selection of many-strain pathogens. *Proceedings of the National Academy of Sciences of the United States of America*. 2002 Dec;99(26):17209–14.

[41] Girvan M, Callaway D, Newman M, Strogatz S. Simple model of epidemics with pathogen mutation. *Physical Review E*. 2002 Mar;65(3):031915.

[42] Ferguson NM, Galvani AP, Bush RM. Ecological and immunological determinants of influenza evolution. *Nature*. 2003;422(6930):428–33.

[43] Tria F, Lässig M, Peliti L, Franz S. A minimal stochastic model for influenza evolution. *Journal of Statistical Mechanics: Theory and Experiment*. 2005 Jul;2005(07):P07008–P07008.

[44] Andreasen V, Sasaki A. Shaping the phylogenetic tree of influenza by cross-immunity. *Theoretical population biology*. 2006 Sep;70(2):164–73.

[45] Boni MF, Gog JR, Andreasen V, Feldman MW. Epidemic dynamics and antigenic evolution in a single season of influenza A. *Proceedings Biological sciences / The Royal Society*. 2006 Jun;273(1592):1307–16.

[46] Omori R, Adams B, Sasaki A. Coexistence conditions for strains of influenza with immune cross-reaction. *Journal of theoretical biology*. 2010 Jan;262(1):48–57.

[47] Koelle K, Cobey S, Grenfell B, Pascual M. Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science (New York, NY)*. 2006 Dec;314(5807):1898–903.

[48] Wu JT, Wein LM, Perelson AS. Optimization of influenza vaccine selection. *Operations Research*. 2005;53(3):456–476.

[49] Gupta V, Earl DJ, Deem MW. Quantifying influenza vaccine efficacy and antigenic distance. *Vaccine*. 2006;24(18):3881–3888.

[50] Gross PA, Hermogenes AW, Sacks HS, Lau J, Levandowski RA. The efficacy of influenza vaccine in elderly persons: a meta-analysis and review of the literature. *Annals of Internal medicine*. 1995;123(7):518–527.

[51] Grotto I, Mandel Y, Green M, Varsano N, Gdalevich M, Ashkenazi I, Shemer J. Influenza

vaccine efficacy in young, healthy adults. *Clinical Infectious Diseases*. 1998;26(4):913–917.

[52] Vu T, Farish S, Jenkins M, Kelly H. A meta-analysis of effectiveness of influenza vaccine in persons aged 65 years and over living in the community. *Vaccine*. 2002;20(13):1831–1836.

[53] Osterholm MT, Kelley NS, Sommer A, Belongia EA. Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis. *The Lancet infectious diseases*. 2012;12(1):36–44.

[54] Myers MF, Rogers D, Cox J, Flahault A, Hay S. Forecasting disease risk for increased epidemic preparedness in public health. *Advances in Parasitology*. 2000;47:309–330.

[55] Koelle K, Rasmussen DA. Influenza: Prediction is worth a shot. *Nature*. 2014 Mar;507(7490):47–8.

[56] Choi K, Thacker SB. An evaluation of influenza mortality surveillance, 1962–1979 I. Time series forecasts of expected pneumonia and influenza deaths. *American journal of epidemiology*. 1981;113(3):215–226.

[57] Stroup DF, Thacker SB, Herndon JL. Application of multiple time series analysis to the estimation of pneumonia and influenza mortality by age 1962–1983. *Statistics in medicine*. 1988;7(10):1045–1059.

[58] McNown R, Rogers A. Forecasting mortality: A parameterized time series approach. *Demography*. 1989;26(4):645–660.

[59] Lee RD, Carter LR. Modeling and forecasting US mortality. *Journal of the American statistical association*. 1992;87(419):659–671.

[60] Quenel P, Dab W. Influenza A and B epidemic criteria based on time-series analysis of health services surveillance data. *European Journal of Epidemiology*. 1998;14(3):275–285.

[61] Cardinal M, Roy R, Lambert J. On the application of integer-valued time series models for the analysis of disease incidence. *Statistics in Medicine*. 1999;18(15):2025–2039.

[62] Viboud C, Boëlle PY, Carrat F, Valleron AJ, Flahault A. Prediction of the spread of influenza epidemics by the method of analogues. *American Journal of Epidemiology*. 2003;158(10):996–1006.

[63] Soebiyanto RP, Adimi F, Kiang RK. Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. *PloS one*. 2010;5(3):e9450.

[64] Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*. 2012 nov;109(50):20425–20430.

[65] Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-time influenza forecasts during the 2012–2013 season. *Nature Communications*. 2013 dec;4.

[66] Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In: AMIA Annual Symposium Proceedings. vol. 2006. American Medical Informatics Association; 2006. p. 244.

[67] Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RA. Using internet searches for influenza surveillance. *Clinical infectious diseases*. 2008;47(11):1443–1448.

[68] Hulth A, Rydevik G, Linde A. Web queries as a source for syndromic surveillance. *PloS one*. 2009;4(2):e4378.

[69] Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457(7232):1012–1014.

[70] Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, Rothman RE. Influenza Forecasting with Google Flu Trends. *PLoS ONE*. 2013 feb;8(2):e56176.

[71] Araz OM, Bentley D, Muelleman RL. Using Google Flu Trends data in forecasting influenza-like–illness related ED visits in Omaha, Nebraska. *The American journal of emergency medicine*. 2014;32(9):1016–1023.

[72] Santillana M, Zhang DW, Althouse BM, Ayers JW. What can digital disease detection learn from (an external revision to) Google Flu Trends? *American journal of preventive medicine*. 2014;47(3):341–347.

[73] Preis T, Moat HS. Adaptive nowcasting of influenza outbreaks using Google searches. *Royal Society open science*. 2014;1(2):140095.

[74] Yang S, Santillana M, Kou S. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences*. 2015;112(47):14473–14478.

[75] Ritterman J, Osborne M, Klein E. Using prediction markets and Twitter to predict a swine flu pandemic. In: 1st international workshop on mining social media. vol. 9. ac. uk/miles/papers/swine09. pdf (accessed 26 August 2015); 2009. p. 9–17.

[76] Culotta A. Towards detecting influenza epidemics by analyzing Twitter messages. In: Proceedings of the first workshop on social media analytics. ACM; 2010. p. 115–122.

[77] Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS one*. 2011;6(5):e19467.

[78] Achrekar H, Gandhe A, Lazarus R, Yu SH, Liu B. Predicting flu trends using twitter data. In: Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on. IEEE; 2011. p. 702–707.

[79] Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one*. 2013;8(12):e83672.

[80] Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. *PLoS currents*. 2014;6.

[81] Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with Wikipedia. *PLoS Comput Biol*. 2014;10(11):e1003892.

[82] McIver DJ, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. *PLoS Comput Biol*. 2014;10(4):e1003581.

[83] Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A,

Del Valle SY. Forecasting the 2013–2014 influenza season using Wikipedia. *PLoS Comput Biol*. 2015;11(5):e1004239.

[84] Dredze M, Cheng R, Paul MJ, Broniatowski D. HealthTweets.org: a platform for public health surveillance using twitter. In: AAAI Conference on Artificial Intelligence; 2014. p. 1–2.

[85] Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput Biol*. 2015;11(10):e1004513.

[86] Biggerstaff M, Alper D, Dredze M, Fox S, Fung ICH, Hickmann KS, Lewis B, Rosenfeld R, Shaman J, Tsou MH, Velardi P, Vespignani A, Finelli L. Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge. *BMC Infectious Diseases*. 2016;16(1):1–10.

[87] CDC Competition Encourages Use of Social Media to Predict Flu. Centers for Disease Control and Prevention; 2013. Available from: `http://www.cdc.gov/flu/news/predict-flu-challenge.htm`. Visited on: 2016-01-18.

[88] Flu Activity Forecasting Website Launched. Centers for Disease Control and Prevention; 2016. Available from: `http://www.cdc.gov/flu/news/flu-forecast-website-launched.htm`. Visited on: 2016-01-19.

[89] DARPA Forecasting Chikungunya Challenge. InnoCentive; 2014. Available from: `https://www.innocentive.com/ar/challenge/9933617?cc=DARPApress&utm_source=DARPA&utm_campaign=9933617&utm_medium=press`. Visited on: 2016-01-18.

[90] Chikungunya threat inspires new DARPA challenge. Science; 2014. Available from: `http://www.sciencemag.org/news/2014/08/chikungunya-threat-inspires-new-darpa-challenge`. Visited on: 2016-01-18.

[91] Chretien JP, Swedlow D, Eckstrand I, George D, Johansson M, Huffman R, Hebbeler A. Advancing Epidemic Prediction and Forecasting: A New US Government Initiative. *Online Journal of Public Health Informatics*. 2015;7(1).

[92] Ebola, Zika modelers aim to inform policy decisions. National Institutes of Health; 2016. Available from: `http://www.fic.nih.gov/News/GlobalHealthMatters/march-april-2016/Pages/disease-modeling-informs-health-policy.aspx`. Visited on: 2016-03-16.

[93] Coiffier J. Fundamentals of Numerical Weather Prediction. Cambridge University Press; 2012.

[94] Klein LR. An Introduction to Econometric Forecasting and Forecasting Models (The Wharton econometric studies series). Lexington Books; 1980.

[95] Chretien JP, George D, Shaman J, Chitale RA, McKenzie FE. Influenza Forecasting in Human Populations: A Scoping Review. *PLoS ONE*. 2014 apr;9(4):e94130.

[96] Nsoesie EO, Brownstein JS, Ramakrishnan N, Marathe MV. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza Other Respi Viruses*. 2013 dec;8(3):309–316.

[97] Racloz V, Ramsey R, Tong S, Hu W. Surveillance of Dengue Fever Virus: A Review of Epidemiological Models and Early Warning Systems. *PLoS Negl Trop Dis*. 2012 may;6(5):e1648.

[98] Galton F. Vox populi (The wisdom of crowds). *Nature*. 1907;75:450–51.

[99] Doswell CA. Weather Forecasting by Humans—Heuristics and Decision Making. *Wea Forecasting*. 2004 dec;19(6):1115–1126.

[100] Surowiecki J. The wisdom of crowds. Anchor; 2005.

[101] Bonabeau E. Decisions 2.0: The power of collective intelligence. *MIT Sloan management review*. 2009;50(2):45–52.

[102] Haykin S. Neural Networks: A Comprehensive Foundation. Prentice Hall PTR; 1994.

[103] Dietterich TG. Ensemble methods in machine learning. In: Multiple classifier systems. Springer; 2000. p. 1–15.

[104] Polgreen PM, Nelson FD, Neumann GR, Weinstein RA. Use of Prediction Markets to Forecast Infectious Disease Activity. *Clinical Infectious Diseases*. 2007 jan;44(2):272–279.

[105] Smolinski MS, Crawley AW, Baltrusaitis K, Chunara R, Olsen JM, Wójcik O, Santillana M, Nguyen A, Brownstein JS. Flu Near You: Crowdsourced Symptom Reporting Spanning 2 Influenza Seasons. *American journal of public health*. 2015;105(10):2124–2130.

[106] Chunara R, Aman S, Smolinski M, Brownstein JS. Flu near you: an online self-reported influenza surveillance system in the USA. *Online Journal of Public Health Informatics*. 2013;5(1).

[107] Viboud C, Nelson MI, Tan Y, Holmes EC. Contrasting the epidemiological and evolutionary dynamics of influenza spatial transmission. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. 2013;368(1614):20120199.

[108] Farrow DC, Burke DS, Rosenfeld R. Computational Characterization of Transient Strain-Transcending Immunity against Influenza A. *PLoS ONE*. 2015 05;10(5):e0125047.

[109] Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, McCauley JW, Russell CA, Smith DJ, Rambaut A. Integrating influenza antigenic dynamics with molecular evolution. *Elife*. 2014;3:e01914.

[110] Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. *Genetics*. 2002;162(4):2025–2035.

[111] A C++ implementation of the FGB2003 model of influenza transmission and evolution. GitHub; 2014. Available from: `https://github.com/undefx/FM_Cpp`. Visited on: 2016-02-29.

[112] A Java implementation of the FGB2003 model of influenza transmission and evolution. GitHub; 2014. Available from: `https://github.com/undefx/FM_Java`. Visited

on: 2016-02-29.

[113] Abu-Raddad LJ, Ferguson NM. The impact of cross-immunity, mutation and stochastic extinction on pathogen diversity. *Proceedings Biological sciences / The Royal Society*. 2004 Dec;271(1556):2431–8.

[114] Glezen WP, Couch RB, MacLean RA, Payne A, Baird JN, Vallbona C, Tristan M, Byrd N. Interpandemic influenza in the Houston area, 1974–76. *New England Journal of Medicine*. 1978;298(11):587–592.

[115] Truscott J, Fraser C, Cauchemez S, Meeyai A, Hinsley W, Donnelly CA, Ghani A, Ferguson N. Essential epidemiological mechanisms underpinning the transmission dynamics of seasonal influenza. *Journal of The Royal Society Interface*. 2011;p. rsif20110309.

[116] Chunara R, Goldstein E, Patterson-Lomba O, Brownstein JS. Estimating influenza attack rates in the United States using a participatory cohort. *Scientific reports*. 2015;5.

[117] Flahault A, Dias-Ferrao V, Chaberty P, Esteves K, Valleron A, Lavanchy D. FluNet as a tool for global monitoring of influenza on the Web. *JAMA : the journal of the American Medical Association*. 1998 Oct;280(15):1330–2.

[118] Cauchemez S, Valleron AJ, Boëlle PY, Flahault A, Ferguson NM. Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature*. 2008 Apr;452(7188):750–4.

[119] Bhatt S, Holmes EC, Pybus OG. The genomic rate of molecular adaptation of the human influenza A virus. *Molecular biology and evolution*. 2011 Sep;28(9):2443–51.

[120] Zinder D, Bedford T, Gupta S, Pascual M. The roles of competition and mutation in shaping antigenic and genetic diversity in influenza. *PLoS pathogens*. 2013 Jan;9(1):e1003104.

[121] Mahalanobis PC. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*. 1936;2:49–55.

[122] Life expectancy. World Health Organization; 2016. Available from: `http://www.who.int/gho/mortality_burden_disease/life_tables/en/`. Visited on: 2016-03-01.

[123] The World Factbook—Median Age. Central Intelligence Agency; 2015. Available from: `https://www.cia.gov/library/publications/the-world-factbook/fields/2177.html`. Visited on: 2016-03-01.

[124] Thomas PG, Keating R, Hulse-Post DJ, Doherty PC. Cell-mediated protection in influenza infection. *Emerg Infect Dis*. 2006;12(1).

[125] Kreijtz J, Bodewes R, van Amerongen G, Kuiken T, Fouchier R, Osterhaus A, Rimmelzwaan G. Primary influenza A virus infection induces cross-protective immunity against a lethal infection with a heterosubtypic virus strain in mice. *Vaccine*. 2007;25(4):612–620.

[126] Kreijtz J, Bodewes R, van den Brand J, de Mutsert G, Baas C, van Amerongen G, Fouchier R, Osterhaus A, Rimmelzwaan G. Infection of mice with a human influenza A/H3N2 virus induces protective immunity against lethal infection with influenza A/H5N1 virus.

*Vaccine*. 2009;27(36):4983–4989.

[127] Bodewes R, Kreijtz JHCM, van Amerongen G, Geelhoed-Mieras MM, Verburgh RJ, Heldens JGM, Bedwell J, van den Brand JMA, Kuiken T, van Baalen CA, Fouchier RAM, Osterhaus ADME, Rimmelzwaan GF. A Single Immunization with CoVaccine HT-Adjuvanted H5N1 Influenza Virus Vaccine Induces Protective Cellular and Humoral Immune Responses in Ferrets. *Journal of Virology*. 2010;84(16):7943–7952.

[128] Cheng X, Zengel JR, Suguitan AL, Xu Q, Wang W, Lin J, Jin H. Evaluation of the humoral and cellular immune responses elicited by the live attenuated and inactivated influenza vaccines and their roles in heterologous protection in ferrets. *Journal of Infectious Diseases*. 2013;208(4):594–602.

[129] Towers S, Feng Z, Hupert N. Short-term heterologous immunity after severe influenza A outbreaks. *arXiv preprint arXiv:10073017*. 2010;.

[130] Lambkin R, McLain L, Jones SE, Aldridge SL, Dimmock NJ. Neutralization escape mutants of type A influenza virus are readily selected by antisera from mice immunized with whole virus: a possible mechanism for antigenic drift. *Journal of General Virology*. 1994 Dec;75(12):3493–3502.

[131] Koelle K, Kamradt M, Pascual M. Understanding the dynamics of rapidly evolving pathogens through modeling the tempo of antigenic change: influenza as a case study. *Epidemics*. 2009 Jun;1(2):129–37.

[132] Wheaton WD, Cajka JC, Chasteen BM, Wagener DK, Cooley PC, Ganapathi L, Roberts DJ, Allpress JL. Synthesized population databases: A US geospatial database for agent-based models. *Methods report (RTI Press)*. 2009;2009(10):905.

[133] Fonville JM, Wilks SH, James SL, Fox A, Ventresca M, Aban M, Xue L, Jones TC, Le NMH, Pham QT, Tran ND, Wong Y, Mosterin A, Katzelnick LC, Labonte D, Le TT, van der Net G, Skepner E, Russell CA, Kaplan TD, Rimmelzwaan GF, Masurel N, de Jong JC, Palache A, Beyer WEP, Le QM, Nguyen TH, Wertheim HFL, Hurt AC, Osterhaus ADME, Barr IG, Fouchier RAM, Horby PW, Smith DJ. Antibody landscapes after influenza virus infection or vaccination. *Science*. 2014;346(6212):996–1000.

[134] Kucharski AJ, Lessler J, Read JM, Zhu H, Jiang CQ, Guan Y, Cummings DaT, Riley S. Estimating the Life Course of Influenza A(H3N2) Antibody Responses from Cross-Sectional Data. *PLoS biology*. 2015 Mar;13(3):e1002082.

[135] Farrow DC, Rosenfeld R. Multiple Resolution Nowcasting of Influenza through Sensor Fusion; 2016. *Manuscript in preparation*.

[136] Brooks FP Jr. No Silver Bullet Essence and Accidents of Software Engineering. *Computer*. 1987 apr;20(4):10–19.

[137] Kalman RE. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*. 1960;82(1):35–45.

[138] Petersen KB, Pedersen MS. The matrix cookbook; 2012. A collection of matrix identities.

[139] Welling M. The kalman filter; 2010. Lecture Note.

[140] Brown RG, Hwang PY. Introduction to random signals and applied Kalman filtering: with

MATLAB exercises and solutions. New York: Wiley; 1997.

[141] Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. *Science*. 2014;343(14 March).

[142] Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PloS one*. 2011;6(8):e23610.

[143] Martin LJ, Xu B, Yasui Y. Improving Google flu trends estimates for the United States through transformation. *PloS one*. 2014;9(12):e109209.

[144] Google Flu Trends. Google; 2015. Available from: `https://www.google.org/flutrends/about/`. Visited on: 2016-03-11.

[145] Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases*. 2009;49(10):1557–1564.

[146] Seifter A, Schwarzwalder A, Geis K, Aucott J. The utility of "Google Trends" for epidemiological research: Lyme disease as an example. *Geospatial health*. 2010;4(2):135–137.

[147] Google Trends. Google; 2016. Available from: `https://www.google.com/trends/`. Visited on: 2016-03-11.

[148] Lampos V, De Bie T, Cristianini N. Flu detector-tracking epidemics on Twitter. In: Machine Learning and Knowledge Discovery in Databases. Springer; 2010. p. 599–602.

[149] Aramaki E, Maskawa S, Morita M. Twitter catches the flu: detecting influenza epidemics using Twitter. In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics; 2011. p. 1568–1576.

[150] HealthTweets. Johns Hopkins Social Media and Health Research Group; 2016. Available from: `http://www.healthtweets.org/`. Visited on: 2016-03-12.

[151] The Delphi Epidemiological Data API. The Delphi Group at Carnegie Mellon University; 2016. Available from: `https://github.com/undefx/delphi-epidata`. Visited on: 2016-03-12.

[152] Greenspan J. Preparing for ILINet 2.0. *Online journal of public health informatics*. 2015;7(1).

[153] Flu Near You. Flu Near You; 2015. Available from: `https://flunearyou.org/`. Visited on: 2016-03-12.

[154] Chunara R, Aman S, Smolinski M, Brownstein JS. Flu near you: an online self-reported influenza surveillance system in the USA. *Online Journal of Public Health Informatics*. 2013;5(1).

[155] Crawley AW, Wojcik O, Olsen J, Brownstein J, Smolinski M. Flu Near You: Comparing Crowdsourced Reports of Influenza-like Illness to the CDC Outpatient Influenza-like Illness Surveillance Network, October 2012 to March 2014. In: 2014 CSTE Annual Conference. Cste; 2014. p. 1.

[156] Smolinski MS, Crawley AW, Baltrusaitis K, Chunara R, Olsen JM, Wójcik O, Santillana M, Nguyen A, Brownstein JS. Flu Near You: Crowdsourced Symptom Reporting Span-

ning 2 Influenza Seasons. *American journal of public health*. 2015;105(10):2124–2130.

[157] Delphi Epicast - Influenza. DELPHI; Carnegie Mellon University; 2015. Available from: `http://epicast.org/`. Visited on: 2016-01-22.

[158] Nelder JA, Mead R. A simplex method for function minimization. *The computer journal*. 1965;7(4):308–313.

[159] Bickel PJ, Levina E. Covariance regularization by thresholding. *The Annals of Statistics*. 2008;p. 2577–2604.

[160] Öllerer V, Croux C. Robust high-dimensional precision matrix estimation. In: Modern Nonparametric, Robust and Multivariate Methods. Springer; 2015. p. 325–350.

[161] Bien J, Tibshirani RJ. Sparse estimation of a covariance matrix. *Biometrika*. 2011;98(4):807–820.

[162] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 2008;9(3):432–441.

[163] Bickel PJ, Levina E. Regularized estimation of large covariance matrices. *The Annals of Statistics*. 2008;p. 199–227.

[164] Hsieh CJ, Dhillon IS, Ravikumar PK, Sustik MA. Sparse inverse covariance matrix estimation using quadratic approximation. In: Advances in Neural Information Processing Systems; 2011. p. 2330–2338.

[165] Kolar M, Xing EP. Consistent covariance selection from data with missing values. In: Proceedings of the 29th International Conference on Machine Learning (ICML-12); 2012. p. 551–558.

[166] Touloumis A. Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings. *Computational Statistics & Data Analysis*. 2015;83:251–261.

[167] Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible Modeling of Epidemics with an Empirical Bayes Framework. *PLOS Computational Biology*. 2015 aug;11(8):e1004382.

[168] Farrow DC, Brooks LC, Hyun S, Tibshirani RJ, Burke DS, Rosenfeld R. A Human Baseline for Epidemiological Forecasting; 2016. *Manuscript under consideration*.

[169] Tribus M. Thermostatics and thermodynamics. Center for Advanced Engineering Study, Massachusetts Institute of Technology; 1961.

[170] Tibshirani RJ. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*. 2014;42(1):285–323.

[171] CHIKV Challenge Announces Winners, Progress toward Forecasting the Spread of Infectious Diseases. Defense Advanced Research Projects Agency; 2015. Available from: `http://www.darpa.mil/news-events/2015-05-27`. Visited on: 2016-01-22.

[172] Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C. Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*. 2000;12(1):19–30.

[173] Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: People erroneously avoid

algorithms after seeing them err. *Journal of Experimental Psychology: General*. 2015;144(1):114–126.

[174] Stuart NA, Market PS, Telfeyan B, Lackmann GM, Carey K, Brooks HE, Nietfeld D, Motta BC, Reeves K. The future of humans in an increasingly automated forecast process. *Bulletin of the American Meteorological Society*. 2006;87(11):1497–1502.

[175] Kittur A, Nickerson JV, Bernstein M, Gerber E, Shaw A, Zimmerman J, Lease M, Horton J. The future of crowd work. In: Proceedings of the 2013 conference on Computer supported cooperative work. ACM; 2013. p. 1301–1318.

[176] Michelucci P, Dickinson JL. The power of crowds. *Science*. 2015 dec;351(6268):32–33.

[177] Kahneman D. Thinking, fast and slow. Macmillan; 2011.

[178] Recker M, Pybus OG, Nee S, Gupta S. The generation of influenza outbreaks by a network of host immune responses against a limited set of antigenic types. *Proceedings of the National Academy of Sciences of the United States of America*. 2007 May;104(18):7711–6.

[179] Symeonidis P, Nanopoulos A, Manolopoulos Y. Feature-weighted user model for recommender systems. In: User Modeling 2007. Springer; 2007. p. 97–106.

[180] Li RH, Yu JX, Huang X, Cheng H. Robust Reputation-Based Ranking on Bipartite Rating Networks. In: SDM. vol. 12. SIAM; 2012. p. 612–623.

[181] van Panhuis WG, Hyun S, Blaney K, Marques Jr ET, Coelho GE, Siqueira Jr JB, Tibshirani R, da Silva Jr JB, Rosenfeld R. Risk of dengue for tourists and teams during the World Cup 2014 in Brazil. *PLOS Negl Trop Dis*. 2014;8(7):e3063.

[182] Racine JS. A primer on regression splines. *URL: http://cranr-projectorg/web/packages/crs/vignettes/splineprimerpdf*. 2014;.

[183] Marsh LC, Cormier DR. Spline regression models. vol. 137. Sage; 2001.

[184] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. vol. 1. Springer series in statistics Springer, Berlin; 2001.

[185] Julier SJ, Uhlmann JK. New extension of the Kalman filter to nonlinear systems. In: AeroSense'97. International Society for Optics and Photonics; 1997. p. 182–193.

[186] FilterPy - Kalman filters and other optimal and non-optimal estimation filters in Python. GitHub; 2016. Available from: `https://github.com/rlabbe/filterpy`. Visited on: 2016-06-04.

[187] CDC Competition Encourages Use of Social Media to Predict Flu. GitHub; 2016. Available from: `https://github.com/rlabbe/Kalman-and-Bayesian-Filters-in-Python`. Visited on: 2016-06-04.